METHODS FOR INCREASING MODEL ACCURACY AND SIMULATION TIME SCALES OF BIOLOGICAL PROCESSES WITH MOLECULAR DYNAMICS

BY

KOON HENG IVAN TEO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Oleksii Aksimentiev, Chair
Professor Klaus Schulten, Director of Research
Professor Emad Tajkhorshid
Associate Professor Yann Chemla
Professor Alfred Hubler

ABSTRACT

This dissertation presents three research projects on novel methods in computational biophysics. Each of these projects introduces methodologies to extend the capabilities of molecular dynamics simulations in one way or another. In the first chapter, molecular dynamics simulations and the central role they play in the field of structural biology is introduced to give the reader some background on the common basis of the projects. The second chapter describes the first of these projects, where the molecular dynamics flexible fitting method for refining molecular structures of macromolecules using experimental electron density data is extended to be able to handle high-resolution density data, which are becoming increasingly commonplace. The third chapter focuses on adaptive multilevel splitting, a replica-based sampling technique that was employed in molecular dynamics simulations to measure the rate of drug molecule dissociation, a process that occurs on the order of milliseconds and above, which is out of the reach of typical molecular dynamics simulations. In the final chapter, a kinetic model of diffusion is introduced. This model allows simulation of the diffusion of small molecules in arbitrary potentials, for example, those that characterize the space around and within a membrane protein channel. The adaptive discretization scheme allows simulations between the micro- to millisecond time scales, which are typical of diffusive processes. This collection of projects is a snapshot of the diversity and versatility of current problems in structural biology that can be addressed by molecular dynamics simulations. I hope to instil in the reader a sense of how method development in molecular dynamics will expand the contributions of the field to both scientific and practical pursuits in biology.

To my parents, for their love and support, and to A'isha, for the affection, patience and care that kept me going.

ACKNOWLEDGMENTS

Much of my academic career has been shaped, both directly and indirectly, by my adviser Klaus Schulten. In our many meetings, I learnt not only how to think about scientific problems, but also how to communicate ideas coherently and convincingly to technical and non-technical audiences. The excellent environment, infrastructure, and community of developers and scientists that Klaus has maintained over the years made it possible for me to bring my research to fruition while developing my own skills.

In the interest of brevity, I would like to acknowledge, out of the numerous interactions I have had with my colleagues, my collaborations with Abhishek Singharoy, Ryan McGreevy, John Stone, and Christopher Mayne. Without their scientific and technical expertise, much of what had been accomplished in this thesis would not have come to pass. I am very fortunate indeed to have been a part of the Theoretical and Computational Biophysics Group.

While completing this thesis has been a difficult undertaking, those close to me have also had to shoulder the burden. For their continued support for my endeavors over the last decade from all the way across the globe, I am grateful to my parents, Robin Teo and Seok Ngoh Choon, and my siblings, KC and Winnie. My girlfriend, Liesel Hess, has always been so patient and encouraging over the years while waiting for my life after graduate school to begin. I would like to share my happiness at the completion of my work in graduate school with Liesel, my family, and everyone who has aided and supported me in one way or another.

TABLE OF CONTENTS

PUBLICATIONS

This thesis has led to the following publications:

CHAPTER 3

A. Singharoy, I. Teo, R. McGreevy (equal credits to preceding authors), J. E. Stone, J. Zhao, and K. Schulten, "Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps," *eLIFE*, **5** (2016), e16105.

CHAPTER 4

I. Teo, C. G. Mayne, K. Schulten, and T. Lelièvre, "Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time," *J. Chem. Theor. Comp.*, **12**(6) (2016), 2983–2989.

D. Aristoff, T. Lelièvre, C. G. Mayne, and I. Teo, "Adaptive multilevel splitting in molecular dynamics simulations," *ESAIM Proc. Surv.*, **48** (2015), 215–225.

CHAPTER 5

I. Teo and K. Schulten, "A computational kinetic model of diffusion for molecular systems," *J. Chem. Phys.*, **139** (2013), 121929.

J. C. Gumbart, I. Teo, B. Roux, and K. Schulten, "Reconciling the roles of kinetic and thermodynamic factors in membrane-protein insertion," *J. Am. Chem. Soc.*, **135**(6) (2013), 2291–2297.

LIST OF ABBREVIATIONS

AMS	Adaptive Multilevel Splitting
BD	Brownian Dynamics
cMDFF	Cascade Molecular Dynamics Flexible Fitting
CHARMM	Chemistry at HARvard Macromolecular Mechanics (name of MD force field)
CPU	Central Processing Unit
ecMscS	E. Coli Mechanosensitive Channel of Small Conductance
EM	cryo-Electron Microscopy
FSC	Fourier Shell Coefficient
GBIS	Generalized Born Implicit Solvent
GCC	Global Cross-correlation Coefficient
GPU	Graphics Processing Unit
iFSC	integrated Fourier Shell Coefficient
LCC	Local Cross-correlation Coefficient
MD	Molecular Dynamics
MDFF	Molecular Dynamics Flexible Fitting
MscS	Mechanosensitive Channel of Small Conductance
MSM	Markov State Model
NAMD	NAnoscale Molecular Dynamics (MD simulation program)
NPT	Isothermal-isobaric ensemble (constant temperature and pressure simulation setup)

NVT	Canonical ensemble (constant temperature and volume simulation setup)
NVE	Microcanonical ensemble (constant energy simulation setup)
PMF	Potential of Mean Force
polyAla	poly-Alanine
polyGln	poly-Glutamine
polyLeu	poly-Leucine
ReMD	Replica Exchange Molecular Dynamics
ReMDFF	Replica Exchange Molecular Dynamics Flexible Fitting
RMSD	Root-Mean-Square Deviation
RMSF	Root-Mean-Square Fluctuation
SA	Signal Anchor
TMD	Targeted Molecular Dynamics
TRPV1	Transient Receptor Potential Cation Channel, subfamily V, member 1
VMD	Visual Molecular Dynamics (molecular structure modelling and analysis program)
WHAM	Weight Histogram Analysis Method

CHAPTER 1

INTRODUCTION

Molecular dynamics (MD) is, by far, the most realistic classical simulation model of molecular systems. From its humble beginnings as simulations of elastic hard spheres [1, 2], and actual physical models made of rubber balls and rods [3], MD has come a long way since the late 1950s. Its numerous successes culminated in the 2013 Nobel Prize in Chemistry awarded to Karplus, Levitt, and Warshel [4] "for the development of multiscale models for complex chemical systems". Nevertheless, MD continues to face difficult methodological challenges that are actively being tackled by many to this day.

MD's claim to realism stems from the fact that it represents systems at the level of individual atoms. Coarse-grained variants do exist where clumps of atoms are aggregated into large particles or the solvent is represented as a continuum, but it remains a fact that the emphasis is on replicating empirical molecular behavior, which is the domain of force field development [5, 6, 7, 8, 9, 10, 11]. The sheer level of complexity of MD simulations challenges the technology and resources required to run such simulations, as well as the ability to obtain good initial models of the simulated systems, since a simulation beginning in an physiologically irrelevant state may not find its way to a relevant one within the course of the simulation.

The work presented in this thesis addresses a few challenges in the field of MD, namely, the accurate inference of macromolecular structure from imaging data, and the simulation of processes that occur on time scales beyond the typical reach of MD simulations. The structures of macromolecules in simulations are not commonly deduced *ab initio*. Typically, the majority of a structure is obtained from experiments that make use of established techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (EM) to obtain an image, or map, of the molecule. Oftentimes, the map does not have sufficient resolution to immediately specify the locations of individual atoms. Fortunately, there is an abundance of structure refinement techniques [12, 13, 14, 15, 16, 17, 18, 19, 20, 21], to find a "best fit" structure for a given low-resolution map.

Traditionally, the gold standard for precision has been X-ray crystallography, for which structures can typically be unambiguously resolved to less than 3 Å. EM has for the most part stayed in the $>$ 5-Å resolution range, but advances in technology in recent years driven the limit down to as low as 1.8 Å [22]. EM also has the advantages over crystallography of not requiring the difficult process of crystallizing the macromolecule, and being able to image the macromolecule in a physiological state (as opposed to a crystalline state). With the rise of high-resolution EM, structure refinement techniques have also had to adapt to the new maps. Chapter 3 deals with the modification of the well-known molecular dynamics flexible fitting (MDFF) algorithm to fit macromolecular structures into a high-resolution EM map, a task which it was previously able to consistently succeed at only with low-resolution maps. The modified techniques, called cascade MDFF and resolution-exchange MDFF, are described in detail, and validated using exemplary biological cases. In addition, this thesis also proposes the use of local fluctuations during an MDFF simulation to evaluate not only how well the structure fits the map, but also how well-resolved the map is around particular regions of the structure.

Another challenge is the simulation of long-time-scale processes. In particular, unbinding processes commonly exhibit time scales of seconds, minutes, or even hours. Chapter 4 describes the application of adaptive multilevel splitting (AMS) [23, 24], a sampling technique previously employed in Monte Carlo simulations, to simulate rare events. Existing methods for simulating such processes typically require the application of artificial forces or potentials to overcome potential barriers [25, 26, 27, 28], which introduces undesirable bias in the dynamics of the system, or the use of prior knowledge to guide the algorithm, for example by setting milestones along the reaction coordinate [29, 30, 31]. AMS utilizes a replica-based paradigm and a scheme to minimize simulation of non-reactive trajectories (trajectories that do not lead to the occurrence of the rare event being studied), to simulate the process in the absence of external forces (in the sense of selectively applying forces to a subset of atoms in the system) without the need for prior knowledge of reaction milestones. A biological test case, the separation of benzamidine from trypsin, was simulated using AMS and the dissociation rate was shown to be in closer agreement with the experimentally determined rate than in other computational efforts to calculate the same rate.

Single-molecule rare events are not the only processes that challenge the time scale of MD simulations. Bulk diffusive processes, such as ion permeation through a membrane channel, can occur on the micro- to millisecond time scale. Such processes typically come under the purview of Brownian dynamics [32, 33, 34, 35], Monte Carlo methods [36], and Green's

function reaction dynamics [37, 38, 39]. Apart from few exceptions [40], most of these methods do not include a detailed, atomistic description of the diffusion environment, lacking either a particular description of the diffusing species or assuming that diffusion occurs in a constant, or simple potential. Chapter 5 introduces a finite-difference kinetic model of diffusion that allows simulation of diffusion particle trajectories across a grid representing the system, under the influence of an arbitrary atomistically-detailed potential map over the grid derived from a prior MD simulation, with dynamics described by the discretized Smoluchowski equation. This method allows simulations spanning microseconds to be performed within hours of clocktime in a serial implementation. A continuum version of the model was applied to investigate the dynamics of nascent chain insertion into the membrane in the SecY translocon, while a single-particle version was applied to small molecule diffusion through the *E. coli* mechanosensitive channel of small conductance (ecMscS) and shown to reproduce the rates of glutamate and potassium ion conductance obtained from a reference study of the same system.

This thesis represents an incremental effort to improve the accuracy of and extend the capabilities of MD to a greater variety of systems and biological processes. The methods presented, although applied to specific test cases, can be generalized to different systems. AMS can be used for the simulation of any rare event, as long as the event can be characterized by transition between well-defined states along some reaction coordinate. The kinetic diffusion model can be employed beyond bulk permeation processes through membrane channel, as shown in the SecY application. As advances in hardware capabilities and accessibility continue to be made, MD will increasingly become a mainstay in the field of structural biology. Methods such as those presented here will help to generalize the application of MD to a wider variety of problems, many of which are of biomedical significance.

CHAPTER 2

MOLECULAR DYNAMICS PRIMER

The work presented in this thesis is centered around molecular dynamics (MD) [1, 41, 42] simulations. The objective of such simulations is to mimic the dynamics of biomolecules in their physiological environments and on the right time scales as realistically as possible. This chapter provides a brief sketch of the basic concepts underlying MD.

In a typical MD simulation, every atom in a system is represented explicitly. There are variants of MD where groups of atoms are aggregated into coarse-grained particles [43, 44, 45, 46, 47, 48, 49], or solvent atoms are represented implicitly [50, 51, 52, 53, 54, 55] for the sake of reducing computational complexity, but these variants will not be discussed here. In the basic MD algorithm, the dynamics of each atom in the system is described by Newton's Second Law:

$$m_\alpha \ddot{\mathbf{r}}_\alpha = -\frac{\partial}{\partial \mathbf{r}_\alpha} U_{\text{total}}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)\,, \tag{2.1}$$

where m_α and $\mathbf{r}_\alpha$ are the mass and position, respectively, of atom α, $\alpha = 1, 2, \ldots, N$, and N is the total number of atoms in the system. The total potential U_{total} consists of several components responsible for the forces acting on each atom.

In an unbiased simulation, U_{total} is given by the following:

$$U_{\text{total}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{UB}} + U_{\text{dihedral}} + U_{\text{improper}} + U_{\text{vdW}} + U_{\text{Coulomb}}\,. \tag{2.2}$$

The first five terms describe bonded interactions using, with the exception of U_{dihedral}, har-

monic functions:

$$U_{\text{bond}} = \sum_{i\in\{\text{bonds}\}} k_i^{\text{bond}}(l_i - l_{0i})^2 \tag{2.3}$$

$$U_{\text{angle}} = \sum_{i\in\{\text{angles}\}} k_i^{\text{angle}}(\theta_i - \theta_{0i})^2 \tag{2.4}$$

$$U_{\text{UB}} = \sum_{i\in\{\text{UB 3-atoms}\}} k_i^{\text{UB}}(r_i^{(13)} - r_{0i}^{(13)})^2 \tag{2.5}$$

$$U_{\text{dihedral}} = \sum_{i\in\{\text{dihedrals}\}} k_i^{\text{dihe}}(1 + \cos(n_i\phi_i - \phi_{0i})) \tag{2.6}$$

$$U_{\text{improper}} = \sum_{i\in\{\text{impropers}\}} k_i^{\text{impr}}(\xi_i - \xi_{0i})^2\,, \tag{2.7}$$

In order of the expressions listed above, these terms respectively model bond stretching, angle-bending, the Urey-Bradley force, torsional angle rotation, and improper angle rotation. The Urey-Bradley force is included only in certain force fields (see below), such as the CHARMM force field [56]. Note that in some conventions, each term is multiplied by a factor of $\frac{1}{2}$, but here we have absorbed it into the force constants.

The Urey-Bradley force and improper dihedral terms were introduced to render simulations consistent with experimentally determined vibrational frequencies [56]. The Urey-Bradley force term controls the 1,3-distance $r^{(13)}$, the distance between the first and third atoms in a bonded series of three atoms, while the improper dihedral term applies to a pyramidal configuration of atoms through the angle ξ between two planes - one containing the three pyramidal base atoms and the other containing two base atoms and the apex atom. The other quantities used in the potentials include the force constants k, the lengths of bonds, l, angles between bonds, θ, dihedral angles, ϕ, and multiplicities of minima in dihedral potentials, n, and their respective constant parameters labelled by a '0' in the subscript.

The remaining terms in U_{total}, the non-bonded interactions, are given by

$$U_{\text{vdW}} = \sum_{i\,,\,j>i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.8}$$

$$U_{\text{Coulomb}} = \sum_{i\,,\,j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\,, \tag{2.9}$$

representing van der Waal's forces through the Lennard-Jones potential [57] and electrostatic forces. Here, i and j are atom labels, σ's are parameters determined empirically, and q's are

the partial charges of the atoms.

The values of all the parameters listed are contained in self-consistent force fields and determined from *ab initio* calculations or fitting to experimental data, depending on the force field being used. The simulations presented in this thesis employ the CHARMM force field [56, 58]. Other popular force fields used in computational biophysics include AMBER [59] and GROMOS [60].

A MD simulation proceeds through numerical integration of Eq. 2.1, and can be performed using either an NVE (constant energy), NVT (constant volume) or NPT (constant pressure) setup. The choice of setup depends on the system to be simulated. For example, a single water droplet containing a small protein can be approximated as a closed system, making NVE a suitable setup. For a large protein requiring a large volume of water, periodic boundary conditions are typically used to approximate an infinite system. To model such a system realistically one would hold the temperature and pressure constant, so that an NPT setup is required. However, controlling both pressure and temperature can sometimes lead to instabilities, especially when the initial state of the system is not within a potential minimum. Thus, it is common practice to equilibrate the system in the more stable NVT setup, prior to performing a production simulation in NPT. Details of the integrator, electrostatics scheme, as well as the thermostat and barostat algorithms used to control temperature and pressure are beyond the scope of this primer.

CHAPTER 3

MDFF FOR HIGH-RESOLUTION CRYO-ELECTRON MICROSCOPY MAPS[1]

Molecular Dynamics Flexible Fitting (MDFF) is a well-established technique for refining atomic models of macromolecules. The refinement is performed by fitting the atomic structure of a macromolecule to a corresponding density map, obtained usually from cryo-electron microscopy (EM) or low-resolution X-ray crystallography experiments, while staying within the constraints of molecular dynamics force fields. As recent advances in EM techniques dramatically increased the resolution of experimental data, it became necessary for MDFF to adapt to continue providing well-fitted structures. This study describes cascade MDFF (cMDFF) and resolution exchange MDFF (ReMDFF), two modified MDFF procedures to meet the challenge posed by high-resolution structures, as well as a novel fluctuation-based protocol for evaluating the quality of fit and quality of model.

3.1 Introduction

Structural biology is built upon the foundation of biomolecular structures that specify the position of every atom within a given biomolecule. These structures are usually obtained experimentally through imaging techniques applied to samples of biomolecules, producing 3D density maps reflecting the positions of atoms within the biomolecules. The most common imaging techniques include X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy (EM). The structure of a given biomolecule can be elucidated directly if the corresponding map is of a sufficiently high resolution, as is often the case with X-ray crystallography. While cryo-EM is typically lower in resolution than X-ray crystallography, it does not suffer from a few difficulties associated with crystallography, namely, the arduous task of preparing well-ordered crystals of macromolecules [61], and the more fundamental problem of capturing these molecules in unphysiological states as a result of

[1]The presented in this chapter has been published in A. Singharoy, I. Teo, R. McGreevy (equal credits to preceding authors), J. E. Stone, J. Zhao, and K. Schulten, *eLIFE*, **5** (2016), e16105. .

crystal contacts [62].

Historically, computational methods were required to bridge the resolution gap between crystallography and cryo-EM to produce atomic-resolution models of biomolecular complexes. Various real-space refinement methods that combine crystallographic structures and cryo-EM densities for structure determination have been developed, including DireX [14], Flex-EM [15], Rosetta [20], FRODA [16], Phenix real space refinement [17], and the family of flexible fitting methods [12, 13, 63, 64, 15, 65, 18, 19], including Molecular Dynamics Flexible Fitting (MDFF) [66, 67, 21].

MDFF, in particular, has proven to be an extremely successful refinement method as evidenced by its numerous applications [68, 21] ranging from the intricate ribosomal machinery [69, 70, 71, 72] to a host of non-enveloped viruses [73]. So far this success has been limited to structure determination from typically low-resolution cryo-EM maps in the $7-25$ Å range which, indeed, represented the state-of-the-art at the time of MDFF's inception [66]. However, seminal advances in detection hardware and programs over the past three years [74, 75] have enabled now the routine availability of high-resolution (< 5 Å) EM maps for a range of biological systems including ion channels [76], enzymes [77, 78], membrane fusion machinery [79] and key functional components of the ribosome [80, 81].

In MDFF, an initial atomic structure, obtained either through *de novo* modelling or another imaging technique, is simulated using molecular dynamics (MD) with its atoms coupled to forces derived from gradients within the corresponding density map. The overall effect is to pull the molecule's atoms into regions of high density within the map, so that the molecule as a whole conforms to the shape of the high-density regions. Although more computationally expensive in comparison to other strategies such as *de novo* modelling [82, 83, 20], MDFF has the advantage of being able to fit the atomic structure while obeying the geometrical constraints imposed by the MD force field, so that structural anomalies are minimized.

MDFF can be thought of as a gradient descent on the sum of the usual MD interaction energy terms and the coupling of atoms to map gradients through the MDFF potential. Like gradient descent, MDFF is prone to being trapped in local minima if the objective function, as in the case of total energy, is non-convex. However, the problem of local minima can be avoided if the starting structure is good, i.e. close enough to the "correct" structure, or if the local minima are shallow enough that temperature-induced fluctuations can push the system out of them. However, in the case of high-resolution maps [76, 77], the level of detail allows even individual side chains to be resolved, giving rise to narrow and steep local minima. Under such conditions, traditional MDFF can result in badly fitted structures. Ob-

taining optimal structures would require extremely precise structure building and validation protocols [20].

In order to address the challenge posed by high-resolution maps, a modification of the direct MDFF method [66, 67, 21], called cascade MDFF (cMDFF), is proposed. Given a starting structure and a high-resolution map for fitting, Gaussian blurring is first applied to the map to obtain a series of maps of different and lower resolutions than the original map. MDFF is then employed to sequentially fit the structure to each map, starting with the map of lowest resolution, and progressing to higher resolutions and ending with the original map. This approach allows the system to escape from local minima early in the fitting process when the minima are still wide and shallow. A second modified protocol inspired by the replica exchange method, resolution exchange MDFF (ReMDFF), achieves the same effect as cMDFF but with a shorter amount of compute time and greater degree of automation. In ReMDFF, an ensemble of starting structures is initialized, with each replica being fitted initially to a map of different resolution. At regular intervals, replicas may exchange maps with each other with a probability depending on the difference in total energy between the two replicas. It should be noted that the concept of using multi-resolution maps for fitting has previously been used successfully for crystallographic data [84].

Another issue raised by the advent of high-resolution density maps is the need for local measures of fit. Traditionally, the quality of fit between an atomic structure and a density map is reflected in a cross-correlation coefficient aggregated over the entire structure. However, as maps become more finely detailed, one can begin to ask how much local uncertainty is associated with particular portions of the molecule. The use of a local cross-correlation coefficient may address this issue in terms of quality of fit. On the other hand, quality of fit may be misleading in the case of a map with heterogeneous resolutions. A high local correlation coefficient of a molecule segment within a map region can be as indicative of low local map resolution as it is of good placement of the residue. In the present study, a multitude of fit and quality metrics will be employed for a robust evaluation of the outcomes of cMDFF and ReMDFF.

In the following sections, the direct MDFF method will be introduced in detail, before the cMDFF and ReMDFF methods are described. The next section will describe applications of cMDFF and ReMDFF to fit available structures to 3.2-Å and 3.4-Å resolution maps of β-galactosidase [77] and the TRPV1 channel [76], respectively. The resulting fits were found to be of accuracy greater than that of direct MDFF and comparable to that of Rosetta, even with poor choices of initial structures. The accuracy is evaluated in terms of the quality of

fit comprehensively measured through global and local cross-correlations (GCC and LCC), integrated Fourier shell coefficients (iFSC), and EMRinger scores [85], as well as in terms of quality of model measures like MolProbity [86].

The second part of this chapter proposes the use of spatial fluctuations during simulation, specifically the local root mean square fluctuation (RMSF), to simultaneously evaluate the goodness of fit of the atomic structure and the quality of the map. In particular, it will be demonstrated that for a given structure-map pair, local RMSF values during MDFF simulation correlate with local cross-correlation between the fitted structure and the map. In addition, local RMSF values during an unbiased MD simulation are shown to correlate with the local resolutions of the map regions containing the corresponding residues in the structure. From these observations, RMSF can be interpreted as a conformational ensemble-based indicator of map quality and of structure fit quality.

3.2 Direct MDFF

The basic MDFF method, direct MDFF, requires for input an initial structure and the EM density map that the structure will be fitted to. An MDFF potential map is generated by inverting and scaling the density map, and is subsequently applied to selected atoms within the initial structure in an MD simulation. The structure thus "feels" the EM-derived potential while simultaneously undergoing structural dynamics as described by the usual MD force field.

Let the density associated with the EM map be $\Phi(\mathbf{r})$. Then the MDFF potential map is given by

$$V_{\mathrm{EM}}(\mathbf{r}) = \begin{cases} \zeta \left(\frac{\Phi(\mathbf{r}) - \Phi_{\mathrm{thr}}}{\Phi_{\max} - \Phi_{\mathrm{thr}}} \right) & \text{if } \Phi(\mathbf{r}) \geq \Phi_{\mathrm{thr}} \,, \\ \zeta & \text{if } \Phi(\mathbf{r}) < \Phi_{\mathrm{thr}} \,. \end{cases} \tag{3.1}$$

where ζ is a scaling factor that controls the strength of the coupling of atoms to the MDFF potential, Φ_{thr} is a threshold for disregarding noise, and $\Phi_{\max} = \max(\Phi(\mathbf{r}))$. The potential energy contribution from the MDFF forces is then

$$U_{\mathrm{EM}} = \sum_i w_i V_{\mathrm{EM}}(\mathbf{r}_i) \,, \tag{3.2}$$

where i labels the atoms in the structure that are coupled to the MDFF potential and w_i is an atom type-dependent weight, usually the atomic mass.

During the simulation, the total potential acting on the system is given by

$$U_{\text{total}} = U_{\text{MD}} + U_{\text{EM}} + U_{\text{SS}} \tag{3.3}$$

where U_{MD} is the MD potential energy as provided by MD force fields, e.g., CHARMM, and U_{SS} is a secondary structure restraint potential (see section on Restraints below) that prevents warping of the secondary structure by the potentially strong forces due to U_{EM}. A detailed description of the potentials arising in Eq. 3.3 is given in Trabuco et al [66, 67].

After the MDFF and restraint potentials are created through the MDFF plugin of VMD [87], the initial structure is rigid-body docked, e.g., with Situs [88], into the density map. Prior to simulation, MDFF-specific parameters can be modified and include ζ and the subset of atoms to be coupled to the MDFF potential. The latter typically consists of all non-hydrogen atoms or backbone atoms and ζ is usually set to 0.3. MDFF can be performed in various simulated conditions, including different temperatures and vacuum, membrane, explicit or implicit [55] solvent environments. The choice of parameters and conditions depends on the requirements of each specific case. For example, a highly polar molecule would be more accurately simulated in explicit solvent rather than in vacuum, but the computational cost would be much higher in this case. A check for correct stereoisomeric orientations is also performed on the structure, as described in the following section, so that any structural errors found can be corrected prior to simulation. The MDFF simulation is run until the system has reached stationarity, as determined by RMSD; typical run times range between 1 to 5 nanoseconds.

3.3 Strategies for High Resolution Density Maps

Flexible fitting methods have facilitated structure determination from low-resolution EM maps for more than a decade [12, 13, 63, 64, 15, 65, 18, 19] and continue to be the methods of choice for resolving molecular systems with atomic resolution. MDFF, in particular, has been a front-runner among methods that have facilitated the discovery of some of the most complicated structures in modern day structural biology [89, 90, 91, 71, 73, 72].

The advent of high-resolution EM maps presents a new challenge to MDFF. In an MDFF simulation, the molecule is allowed to reach equilibrium while under the influence of both the MD and the guiding MDFF potentials. Ideally, the equilibrium structure obtained in the simulation represents a global minimum in the total energy, which is dominated by

the MDFF potential map V_{EM}. For maps in the low resolution range (6 to 15 Å), this global minimum is broad, accommodating an ensemble of conformations defined by the overall shape of the molecule [66, 70]. In contrast, at the mid-resolution range of 4 to 6 Å, densities corresponding to the backbones become discernible, and at sub-4 Å resolutions, even sidechains can be resolved. At such high resolutions, V_{EM} features multiple proximal local minima which correspond to recurring spatial patterns within a macromolecule, such as helices aligned in parallel or strands in a β-sheet. The energy barriers separating these local minima are typically about twice as high as those in the case of low-resolution maps. The existence of such potential minima in high-resolution maps exposes MDFF to a long-known weakness of traditional MD-based algorithms, namely entrapment of the fitted structure within undesired local minima instead of reaching the global minimum of V_{EM}. As a result, direct MDFF yields structurally poor or functionally irrelevant models with high-resolution EM maps (Fig. 3.1) [20].

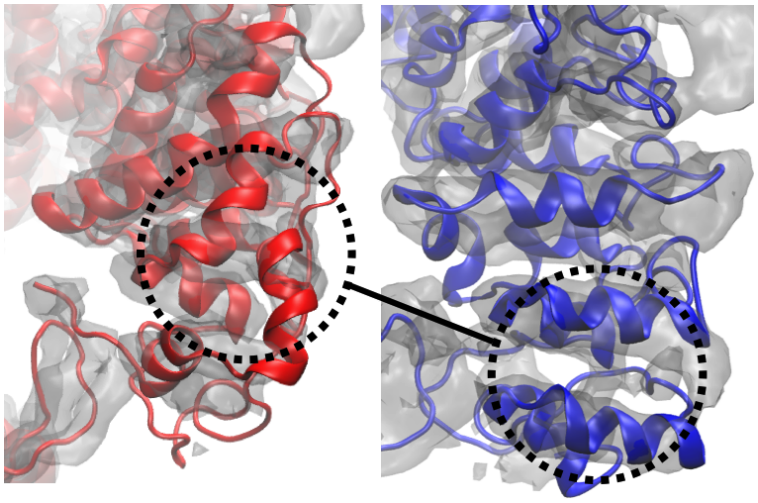

Figure 3.1: Global cross-correlation as a measure of fit. The blue and red structures represent the same region of a segment of TRPV1 that have been fitted differently into the density map shown. The global cross-correlations of the structural region shown in each case are 0.728 (red) and 0.723 (blue). However, the blue structure is clearly better fitted than the red structure, as reflected in RMSDs from the published structure of 6.2 Å (red) and 2.3 Å (blue). Although the case described is an extreme one, it shows that global cross-correlation, as a measure of fit, can be misleading, particularly in regards to local correspondence of residues to the map.

The new variants of MDFF, cMDFF and ReMDFF, were designed to overcome the limitation resulting from local minima, allowing accurate fitting of molecular structures within

sub-5 Å EM maps. These new methods extend the radius of convergence of MDFF to at least 25 Å, fitting models to maps of resolutions as high as 3.2 Å. This radius of convergence is at least twice that reported for Rosetta refinements of the 20S proteasome [20]. Such a broad radius of convergence will allow refinement of extremely poorly guessed initial models with MDFF, as demonstrated in the cases of β-galactosidase and TRPV1 discussed in this chapter.

ReMDFF simulations involving the so-called replica-exchange molecular dynamics method converge quickly using a small number of replicas and are thus amenable to cloud computing applications. Running ReMDFF on the cloud greatly lowers the barrier to usage of the method providing a cost-effective and practical solution to fitting structures to high-resolution cryo-EM densities for researchers who neither own nor can administer their own advanced computer hardware.

3.3.1 Cascade MDFF

In cascade MDFF (cMDFF), the initial structure is sequentially fitted to a series of potential maps of successively higher resolution, with the final potential map being the original one derived from the EM map. Starting with $i = 1$, the ith map in the series is obtained by applying a Gaussian blur of width $\sigma_i \geq 0$ Å to the original potential map, such that σ_i decreases as the structure is fitted in the sequence $i = 1, 2, \ldots, L$, where L is the total number of maps in the series, so that $\sigma_L = 0$ Å. One can intuitively understand cMDFF as fitting the simulated structure to an initially large and ergodic conformational space that is shrinking over the course of the simulation towards the highly corrugated space described by the original MDFF potential map.

The gradual increase in map resolution over the course of the simulations allows the structure to explore a greater conformational space than in direct MDFF. The structure thus avoids entrapment within local minima of the MDFF potential and is accurately fitted to the near-atomic density features of the experimental map.

To mathematically illustrate the cMDFF concept, begin by describing the Gaussian blur process, which produces a potential map V_σ from the potential map V_{EM} through convolution with a normalized Gaussian of specified width σ:

$$V_\sigma(\mathbf{r}) = \int d\mathbf{r}'\, G(\mathbf{r}\,;\,\mathbf{r}', \sigma) V_{\mathrm{EM}}(\mathbf{r}') \quad , \tag{3.4}$$

where $G(\mathbf{r}\,;\,\mathbf{r}',\sigma)$ denotes a normalized Gaussian of width σ centered at $\mathbf{r}'$ and evaluated at $\mathbf{r}$ given by

$$G(\mathbf{r}\,;\,\mathbf{r}',\sigma) = A(\sigma)\exp\left(-\frac{||\mathbf{r}-\mathbf{r}'||^2}{2\sigma^2}\right) \quad , \tag{3.5}$$

$$A(\sigma) = \frac{1}{(2\pi\sigma^2)^{3/2}}\,. \tag{3.6}$$

One can characterize the resolution of $V_{\rm EM}(\mathbf{r})$ explicitly by assuming that it can be written as a sum of Gaussians.

$$V_{\rm EM}(\mathbf{r}) = \sum_n c_n G(\mathbf{r}\,;\,\mathbf{r}'_n,\sigma'_n) \quad , \tag{3.7}$$

where c_n are weighting factors, $\mathbf{r}'_n$ and σ'_n are, respectively, the centers and widths of the component Gaussians. Substituting the above expression into Eq. (3.4) yields

$$V_\sigma(\mathbf{r}) = \int d\mathbf{r}'\, G(\mathbf{r}\,;\,\mathbf{r}',\sigma)\sum_n c_n G(\mathbf{r}'\,;\,\mathbf{r}'_n,\sigma'_n) \tag{3.8}$$

$$= \sum_n c_n A(\sigma)A'_n(\sigma'_n)\int d\mathbf{r}'\exp\left(-\frac{||\mathbf{r}-\mathbf{r}'||^2}{2\sigma^2}-\frac{||\mathbf{r}'-\mathbf{r}'_n||^2}{2\sigma'^2_n}\right) \quad , \tag{3.9}$$

where $A(\sigma)$ and $A'_n(\sigma'_n)$ are the normalizing factors for $G(\mathbf{r}\,;\,\mathbf{r}',\sigma)$ and $G(\mathbf{r}\,;\,\mathbf{r}'_n,\sigma'_n)$, respectively. Evaluation of the above expression gives

$$V_\sigma(\mathbf{r}) = \sum_n C_n \exp\left(-\frac{||\mathbf{r}-\mathbf{r}'_n||^2}{2(\sigma^2+\sigma'^2_n)}\right) \quad , \tag{3.10}$$

$$C_n = c_n A(\sigma)A'_n(\sigma'_n)(2\pi)^{3/2}\left(\frac{\sigma\sigma'_n}{\sqrt{\sigma^2+\sigma'^2_n}}\right)^3 \tag{3.11}$$

$$= c_n\left[2\pi(\sigma^2+\sigma'^2_n)\right]^{-3/2} \quad . \tag{3.12}$$

Note that setting $\sigma = 0$ Å in Eqs. (3.10-3.12) recovers the expression for the initial map in Eq. (3.7), i.e. $V_0(\mathbf{r}) = V(\mathbf{r})$. On the other hand, increasing σ results in an increase in the widths of the component Gaussians, and leads in turn to an increase in the range of the MDFF force $\mathbf{F}_\sigma(\mathbf{r})$, where

$$\mathbf{F}_\sigma(\mathbf{r}) = -\nabla V_\sigma(\mathbf{r}) \tag{3.13}$$

$$= \sum_n \frac{C_n}{\sigma^2+\sigma'^2_n}\exp\left(-\frac{||\mathbf{r}-\mathbf{r}'_n||^2}{2(\sigma^2+\sigma'^2_n)}\right)(\mathbf{r}-\mathbf{r}'_n) \quad . \tag{3.14}$$

Hence, at each stage i of the cMDFF process, the associated map half-width σ_i allows one to tune the characteristic width of the potential map through the half-widths of its component Gaussians $\sqrt{\sigma_n'^2 + \sigma_i^2}$. The initial fitting starts with a large σ_1, corresponding to a diffuse potential which allows much structural mobility, and proceeds along decreasing values of σ_i, corresponding to narrower potentials with steeper gradients, so that the structure is gradually settled into the original potential map, characterized by $\sigma_L = 0$ Å.

In practice, the series of cMDFF maps is generated from the original potential map using VMD's volutil Gaussian blur tool. Optimal values for the first half-width σ_1 and the change in σ_i from one map to the next are case-dependent. Values used in the present study were obtained through trial-and-error. In general, structures far from the ideal conformation benefit from a large σ_1 (> 5 Å) so as to explore a large conformation space. On the other hand, if the original map has a high resolution, small changes in σ_i (< 1 Å) would allow a gradual convergence required to avoid being trapped in local potential minima. In our simulations, the change in σ_i is initially 1 Å. A concrete example is $\sigma_1 = 5$ Å, $\sigma_2 = 4$ Å, $\sigma_3 = 3$ Å, $\sigma_4 = 2$ Å, $\sigma_5 = 1$ Å, $\sigma_6 = 0$ Å. A second trial using changes of 0.5 Å was performed ($\sigma_1 = 5$ Å, $\sigma_2 = 4.5$ Å, $\sigma_3 = 4$ Å,..., $\sigma_{10} = 0.5$ Å, $\sigma_{11} = 0$ Å), and if the resulting structure of the second trial presented a better fit, then the first trial was discarded.

3.3.2 Resolution Exchange MDFF

ReMDFF increases the degree of automation in the cMDFF method. The discussion of the ReMDFF method here is preceded by a description of Replica Exchange MD (ReMD), from which ReMDFF was inspired. ReMD is an advanced sampling method that explores conformational phase space in search of conformational intermediates, which are separated by energy barriers too high to be overcome readily by fixed temperature simulations. Instead of working with a single, fixed MD simulation, ReMD carries out many simulations in parallel, but at different temperatures $T_1 < T_2 < T_3 < \ldots$ where the lowest temperature T_1 is the temperature of actual interest, typically, room temperature. The simulations of several copies of the system, or replicas, run mainly independently, such that ReMD can be easily parallelized over multiple processors, but at regular time points the instantaneous conformations of replicas of neighboring temperatures are compared in terms of energy and the exchange of temperature values between replicas are permitted according to a Metropolis criterion [92]. Under this scheme, the highest temperature replicas overcome the energy barriers between conformational intermediates while the lower temperature replicas seek out

the most favorable local minima. The application of the Metropolis criterion in the protocol guarantees that the conformations of the T_1 replica are Boltzmann-distributed.

ReMDFF extends the concept of ReMD to MDFF by exchanging the resolutions of MDFF potential maps between replicas instead of temperatures as in the case of ReMD. In the ReMDFF method, the replicas are run at the same temperature, but fitted to maps of varying half-widths. At fixed time intervals, the replicas i and j, characterized by atom coordinates $\mathbf{x}_i$ and $\mathbf{x}_j$ and fitting maps of blur widths σ_i and σ_j, are compared energetically and exchanged with Metropolis acceptance probability

$$p(\mathbf{x}_i, \sigma_i, \mathbf{x}_j, \sigma_j) = \min\left(1\,,\, \exp\left(\frac{-E(\mathbf{x}_i, \sigma_j) - E(\mathbf{x}_j, \sigma_i) + E(\mathbf{x}_i, \sigma_i) + E(\mathbf{x}_j, \sigma_j)}{k_B T}\right)\right) \quad , \tag{3.15}$$

where k_B is the Boltzmann constant, $E(\mathbf{x}, \sigma)$ is the instantaneous total energy of the configuration $\mathbf{x}$ within a fitting potential map of blur width σ.

A replica that is stuck within a bad local minimum in a high-resolution map can then escape after exchanging with a replica being fitted to a low-resolution map. On the other hand, a replica within a low-resolution map can be more precisely fitted after exchanging with a replica being fitted to a high-resolution map. The parallelization capabilities of NAMD implemented for ReMD [93] are easily extended to ReMDFF, so that the enhanced sampling achieved translates into extremely fast MDFF convergence.

3.4 Methods

In a proof-of-principle case, various MDFF simulations were applied to carbon dioxide dehydrogenase to illustrate conceptually the advantages of cMDFF and ReMDFF over direct MDFF. Details are presented in MDFF Results. Following which, two test cases were used to evaluate the performance of cMDFF and ReMDFF relative to direct MDFF. Both cases are well-known pioneer instances of high-resolution EM maps - β-galactosidase and TRPV1. The former is a glycoside hydrolase enzyme in solution and the latter is a temperature-sensing membrane channel. Both macromolecules are homotetramers with available atomic structures inferred *de novo* from their respective EM maps. These structures were used as benchmarks for comparison with the results of fitting, and also as precursors for the initial structures for the simulations. The following sections detail the protocols used in the direct MDFF, cMDFF and ReMDFF refinements of the macromolecules in the two test cases.

3.4.1 MD Simulation Setup

Unless otherwise stated, all simulations reported in the present study used the following MD parameters. MDFF simulations were run in vacuum, maintained at 300 K via a Langevin thermostat and employed a time step of 1 fs. Secondary structure restraints were imposed using NAMD's Extra Bonds function to prevent loss of secondary structure due to strong MDFF guiding forces.

MD simulations were prepared using CHARMM-GUI [94] and run under NPT conditions, maintained at 303.15 K temperature and 1 atm pressure using a Langevin thermostat and barostat. All systems were parameterized using the CHARMM36 force field [95, 58]. Structures were solvated in explicit water (TIP3P model) boxes, with at least 15 Å separation between structure and water box boundaries. Particle-mesh Ewald electrostatics was used and the time step was 2 fs. For simulations of the reported structures of β-galactosidase, backbone atoms were held fixed while a minimization over 1000 time steps was performed, followed by 32-ns and 40-ns equilibration for the 3.2-Å and 2.2-Å structures, respectively. Following the equilibration step, production runs of 30 ns were performed for both structures.

In the case of TRPV1, the channel was embedded in a membrane of standard lipid composition POPE, POPC, POPG at ratio 2:1:1. Initial runs involved minimization over 1000 steps and 1-ns equilibration of the lipid tails, with all other atoms fixed. In the following MD run, minimization was performed over 5000 steps and equilibration was performed for 3 ns, with protein backbone atoms held fixed. Finally, the entire system was equilibrated for a further 6.4 ns. During the equilibration, C-terminal residues 752 to 762 were harmonically restrained because a substantial C-terminal segment was missing in the structure.

3.4.2 Fourier Shell Coefficients

Fourier Shell Coefficients (FSCs) can be used as a means of evaluating quality of fit by comparing the degree of similarity between the original map and a simulated map derived from the structure to be evaluated, using the simulated map feature of VMD's MDFF package [66, 67] and the same voxel size as the original map.

In the present study, FSC curves for fitted TRPV1 and β-galactosidase structures were calculated via the FSC operation in SPIDER [96], using a shell width of 0.5 reciprocal space units, and resolution cutoff of half the voxel size. In the case of TRPV1, both the full structure and MDFF-fitted region were evaluated. The latter was obtained by applying a mask of the region around residues 199 to 430 in the fitted structure to the simulated map,

and in the reported structure to the original map.

As a means of summarizing comparisons by FSC, other studies have used "integrated FSCs", a numerical measure obtained by integrating under the FSC plot within a predefined resolution interval. Two integrated FSC measures, corresponding to the intervals 3.4 Å to 10 Å and 5 Å to 10 Å, were obtained in the present study and tabulated in Table 3.1, and Tables 3.5 and 3.7.

3.4.3 Preparation of initial test structures

The first initial structure of β-galactosidase was obtained by subjecting the reported structure [77] to a 4-ns equilibrium MD simulation at a temperature of 300 K. Trajectory frames recorded at 2-ps intervals were evaluated for backbone RMSDs with respect to the reported structure. A frame with an RMSD value of 7.6 Å (Fig. 3.2a) and lowest global cross-correlation with respect to the reported map was picked to be the initial test structure. The structural quality measure of this model is provided in Table 3.2. A second initial structure was prepared by repeating the same protocol but at 1000 K. This structure is also characterized by an RMSD of 7.6 Å but now features a more distorted local structure as measured in terms of increased rotamer and Ramachandran outliers (Table 3.4).

The initial test structure of TRPV1 was also derived from a reported structure [76]. In order to render the disjointed reported structure contiguous for correct structural dynamics during simulation, the missing loop region (residues 503 to 506) was added by hand. Additionally, the substantial ankyrin repeat region (residues 111 to 198) was removed because the corresponding density was missing from the map. For the purpose of testing the robustness of cMDFF and contrasting its performance with that of direct MDFF, the structure was distorted (see Fig. 3.2b) during an interactive MD [97, 98] simulation, subjecting residues 199 to 430 in one subunit's extramembrane domain to a series of transformations, consisting roughly of a polar angle change of 15° toward the cytoplasmic pole followed by an azimuthal rotation of 30°, so that the backbone RMSD of the transformed region was about 22 Å relative to the original structure.

3.4.4 Direct MDFF.

In order to provide a basis for comparison with cMDFF and ReMDFF, direct MDFF simulations were performed for both β-galactosidase and TRPV1. For β-galactosidase, an EM

map of 3.2-Å resolution (EMD-5995 [77]) is available with corresponding *de novo* structure listed as PDB entry 3J7H. Two initial structures were used for fitting. The first was obtained by simulating the *de novo* structure at equilibrium for 4 ns and selecting a trajectory frame in which the backbone RMSD was at a maximum value of 7.6 Å relative to the *de novo* structure, while the second was obtained through a high temperature equilibration simulation and had the same RMSD but poorer secondary structure quality as measured by the proportion of Ramachandran outliers.

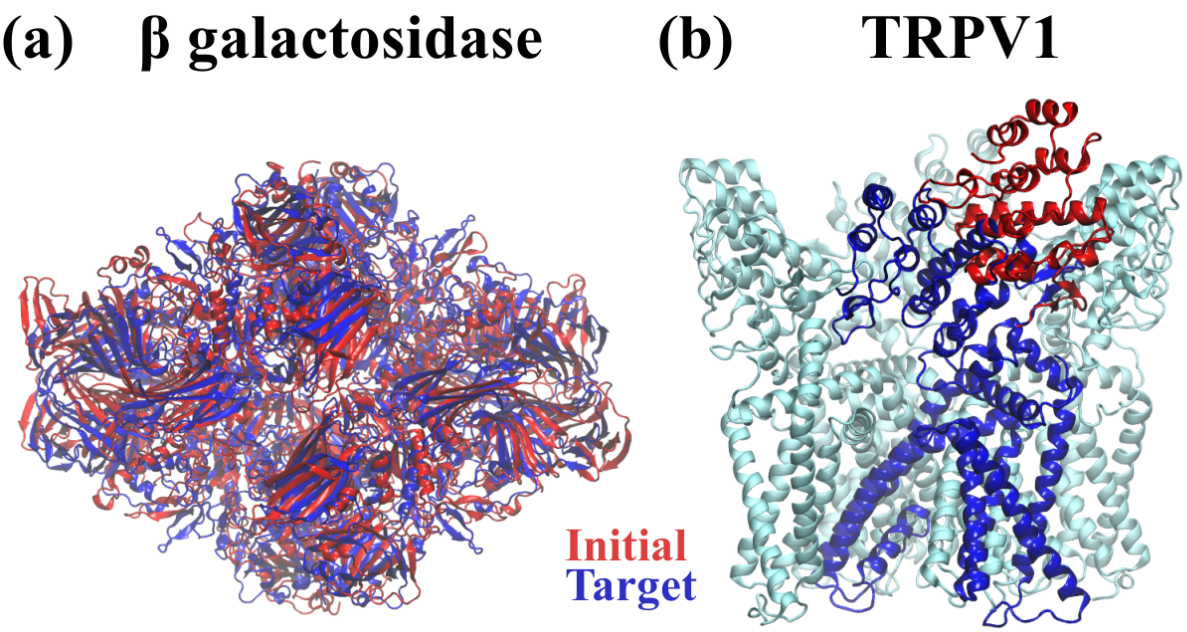

Figure 3.2: Comparison of initial models to target (published) models. For the purpose of testing cMDFF on **(a)** β-galactosidase and **(b)** TRPV1, the published models (blue) were distorted to provide the initial models (red) for fitting. In the case of TRPV1, the distortion was applied to only one subunit.

For TRPV1, the published EM map has a resolution of 3.4 Å (EMD-5778 [76]) and corresponding *de novo* structure listed as PDB entry 3J5P. The latter structure contained a missing loop region (residues 503 to 506) which had to be filled in by hand. Additionally, the substantial ankyrin repeat region (residues 111 to 198) was removed because the corresponding density was missing from the map. To obtain the initial structure for the fitting simulations, the *de novo* structure was distorted (see Fig. 3.2b) during an interactive MD [97, 98] simulation, subjecting residues 199 to 430 in one subunit's extramembrane domain to a series of transformations, consisting roughly of a polar angle change of 15° toward the cytoplasmic pole followed by an azimuthal rotation of 30°, so that the backbone RMSD of

the transformed region was about 22 Å relative to the *de novo* structure. Further details on initial structure preparation for both β-galactosidase and TRPV1 are given in Section 3.4.3.

The initial structures of β-galactosidase and TRPV1 were first minimized over 1000 time steps. Scale factors ζ (Eq. (3.1)) of values 1.0 and 0.3 were employed for β-galactosidase and TRPV1, respectively, to couple all backbone atoms to the respective maps. All other simulation parameters are listed in Section 3.4.1. The resulting structure from each MDFF simulation was then subjected to a final re-refinement step - fitting with a scale factor of 1.0 while the temperature was ramped down from 300 K to 0 K over 30 ps and held at 0 K for an additional 1 ns. This final refinement step was found to improve the fitting of sidechains [99]. With the exception of the ζ values, the pre- and post-MDFF procedures described here were also applied to the cMDFF and ReMDFF simulations.

3.4.5 cMDFF

The cMDFF protocol consists of a series of consecutive MDFF simulations, starting with the map of the lowest resolution, progressing through maps of successively higher resolution, and ending with the map of the original resolution. The duration of each run was long enough (70 ps for β-galactosidase, 100 ps for TRPV1) for the structure to equilibrate within the MDFF potential. To take advantage of the stochastic nature of MDFF simulations, 10 independent cMDFF simulations were performed for each system to be fitted, generating an ensemble of fitted structures. From the ensemble, the best structure was determined by the various quality indicators described in MDFF Results. This structure was then subjected to the final re-refinement step to allow for accurate resolution of sidechains.

For β-galactosidase, cMDFF was initiated with a map blurred with half-width $\sigma_1 = 5$ Å. The subsequent maps were blurred with half-widths decreasing in steps of 0.5 Å, giving $L = 11$ maps in total, including the original. In another set of simulations, we observed that using a larger step size of 1 Å caused the structure to converge to a less well-fitted configuration. For TRPV1, Gaussian blurred maps were generated starting with a half-width of $\sigma_1 = 5$ Å, and decreasing by 1 Å for each subsequent map, thus yielding a series of $L = 6$ maps, including the original.

3.4.6 ReMDFF

ReMDFF was performed on both β-galactosidase and TRPV1 using the same initial structures, simulation parameters and maps as in the cMDFF simulations. 11 and 6 replicas were employed for β-galactosidase and TRPV1 respectively with an exchange trial interval of 1 ps. In each case, the total energy of each replica was monitored and the simulation was run until the energies reached a stationary level. The ReMDFF simulation was found to converge in 0.1 ns for the β-galactosidase refinement, and in 0.02 ns for that of TRPV1. Finally, similar to direct MDFF and cMDFF, the re-refinement step was performed to improve sidechain geometry.

3.4.7 Cross-validation of MDFF-fitted Structures

To demonstrate that the over-fitting does not occur during cMDFF refinements, which is also fairly representative of ReMDFF refinements, the reported *de novo* structures of β-galactosidase and TRPV1 were each fitted to two half-maps (labelled 1 and 2) of the corresponding reported EM map, [77] for β-galactosidase and [76] for TRPV1. Subsequently, simulated maps were created from the fitted structures using VMD's MDFF plugin and resolution settings equivalent to the reported maps. In total, there were two simulated maps, also with labels 1 and 2 corresponding to the half-map from which the fits were obtained, for each protein. FSC plots describing the direct comparison of simulated maps with the corresponding half-maps (e.g. simulated map 1 with half-map 1) as well as the cross comparison of simulated maps with the non-corresponding half-maps (e.g. simulated map 1 with half-map 2) were created. The high degree of similarity between the cross comparisons as well as between cross comparisons and direct comparisons indicate a very low degree of over-fitting. In fact, iFSC values calculated for the plots (see Fig. 3.3) are practically uniform. EMRinger scores for the same sets of comparisons were also calculated. For β-galactosidase, the EMRinger scores were 3.25 for simulated map 1 against half-map 1, 2.97 for simulated map 2 against half-map 2, 2.92 for simulated map 1 against half-map 2, and 2.81 for simulated map 2 against half-map 1; these numbers are fairly comparable to the EMRinger scores with the full maps as presented in Table 3.1. For TRPV1 again, the EMRinger scores were 1.43 for all comparisons. The high degree of similarity between EMRinger scores for the different comparisons corroborate the favorable conclusion drawn from the FSC calculations.

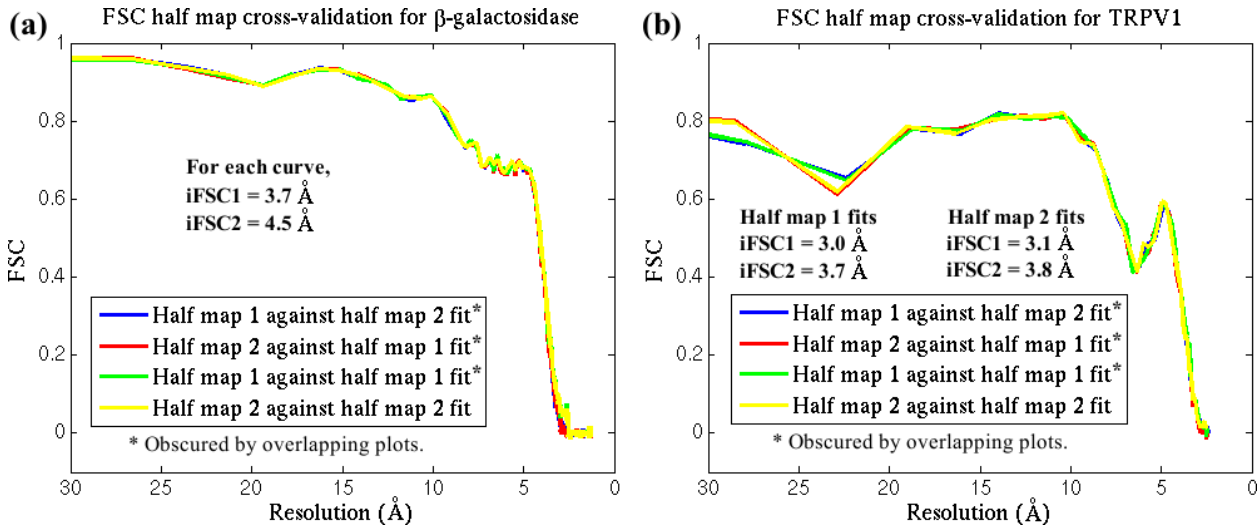

Figure 3.3: FSC cross-validation plots. The reported structures for **(a)** β-galactosidase and **(b)** TRPV1 were each fitted by direct MDFF against two half-maps, labelled 1 and 2, from their respective EM data. Simulated maps were generated from the resulting structures, with labels corresponding to the half-maps used in the fitting. FSC plots of the simulated maps against the half-maps are so similar that they superimpose on one another. In addition, the differences in iFSCs between the various plots are negligible. These results demonstrate that the MDFF method, with the parameters used in the present study, do not overfit the structure.

3.5 MDFF Results

Results from the direct MDFF, cMDFF, and ReMDFF simulations are described in this section. The first section describes a proof-of-principle test case involving a simple molecule and idealized EM map, performed prior to the more realistic cases of β-galactosidase and TRPV1. Results for the latter cases are described thereafter, including both structural evaluations of the refined structures using established methods in the cryo-EM field and efficiency of the ReMDFF protocol on Amazon's cloud computing platform.

3.5.1 Proof of principle: carbon monoxide dehydrogenase

In an initial proof-of-principle computation, direct MDFF, cMDFF and ReMDFF were applied to fit a structure of carbon monoxide dehydrogenase to maps of varying resolutions (see columns 2 and 3 of Fig. 3.4), obtained by Gaussian blurring of a 3-Å synthetic density

map using half-widths (σ) ranging from 5 to 0 Å at constant decrements of 1 Å. Carbon monoxide dehydrogenase exhibits a closed and an open conformation [100]. Both of these conformations have been crystallized, and are reported respectively in chains C and D of the PDB entry 1OAO. For the present demonstration, the closed conformation (1OAO:chain C) was used as the initial structure, while the open one (1OAO:chain D) was the target.

The 3-Å resolution synthetic density map was constructed in Phenix [101], employing phases from the 1OAO structure and the associated diffraction data truncated at 3 Å. This map was then masked about chain D to yield a high-resolution envelope characterizing the open conformation. Assuming that the crystallographic model provides an accurate benchmark, the corresponding map for chain D determined here represents the best possible density data at 3 Å resolution that is experimentally attainable for the open conformation.

Accuracy of the fitting protocols was evaluated by comparing the fitted chain C structures with the crystallographically reported target chain D model. Each row in Fig. 3.4 corresponds to a direct MDFF fit to the map of the stated σ value in the first column. Note that Gaussian blurring of the map lowers potential barriers within the map, as observed in the map cross-sections in the third column of Fig 3.4, so that convergence in RMSD is faster with greater blur half-widths (last column in Fig 3.4). Direct MDFF of the 3 Å synthetic map performed for 2 ns converged to a structure with an RMSD of 7 Å relative to the target model. In sharp contrast, the cMDFF- and ReMDFF-generated structures are within 1.7 Å and 1 Å RMSD of the target (see the inset of Fig. 3.4). Also, note that fitting to the blurred maps produced structures that are around 2 Å RMSD, and subsequent fitting to high-resolution maps, equivalent to the cMDFF protocol, brought the RMSD down to 1.0 Å.

The results demonstrate that the new protocols are capable of attaining well-fit structures where direct MDFF does not. In particular, one can think of the new protocols as extending the radius of convergence to at least 7 Å, rendering the fitting procedures less dependent on the initial configuration of the starting structure.

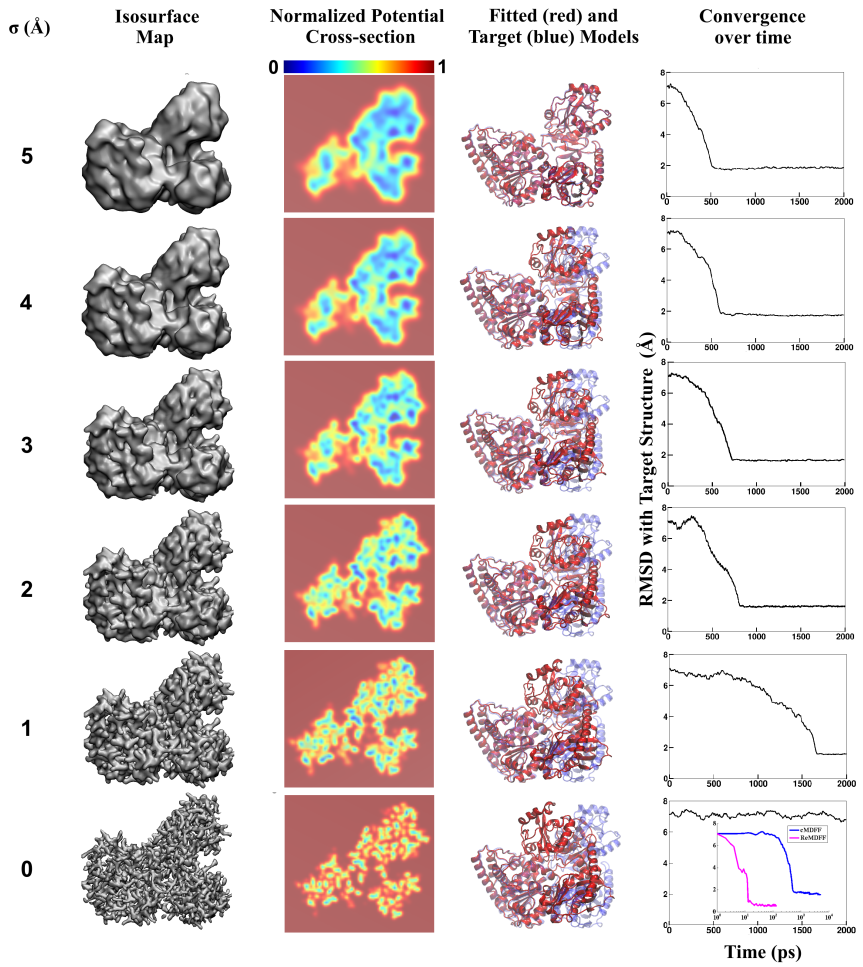

Figure 3.4: Visual summary of advanced MDFF methodology. A graphic table illustrating MDFF refinement of carbon monoxide dehydrogenase using maps of varying resolutions. The maps represent an open conformation while the initial structure was obtained through crystallography of a closed conformation. This initial structure was independently fitted, using direct MDFF, to individual maps, each obtained by applying a Gaussian blur of a different half-width (σ, first column) to the original map. These maps are visualized as isopotential surfaces and cross-sections in the second and third columns, respectively. As σ increases, the amount of contiguous high-density regions increases and the V_{EM} barriers go from high (red) to low (blue). The overall effect is greater freedom for the structure to explore conformational space during fitting. The structure after 500 ps of fitting, shown in red, is superimposed on the known target structure, shown in blue, in the fourth column. The time evolution of RMSD with respect to the target during fitting is shown in the fifth column. Direct fitting to lower resolution maps requires fewer time steps to reach convergence. In particular, RMSD never drops appreciably during fitting to the original map. The inset shows refinements of the same structure by cMDFF and ReMDFF employing the same set of maps. A clear improvement in fit over direct MDFF is apparent, with convergence to within 1.7 Å and 1.0 Å of the target achieved within 1000 and 100 ps for cMDFF and ReMDFF respectively.

3.5.2 A note on fitting metrics

The cross-correlation coefficient calculated over an entire structure, termed global cross-correlation coefficient (GCC), has been a popular indicator of goodness-of-fit of a structure to a corresponding EM density map. However, averaging over the entire structure smears out potentially useful local structure information and in some cases, can be misleading (see Fig. 3.1), since GCC cannot distinguish between correct and wrong assignments of residues to a given map region as long as the residues are equally well fitted.

Local measures of fit allow one to assess every part of the structure individually. In the present study, local cross-correlation coefficients (LCCs) [102] were tracked over the course of the simulations of β-galactosidase and TRPV1 (Fig. 3.5). The improvement in LCC of the majority of residues in each case lends greater confidence in the fitting result. At the same time, residues that have relatively lower LCCs can be identified for further treatment [102, 21].

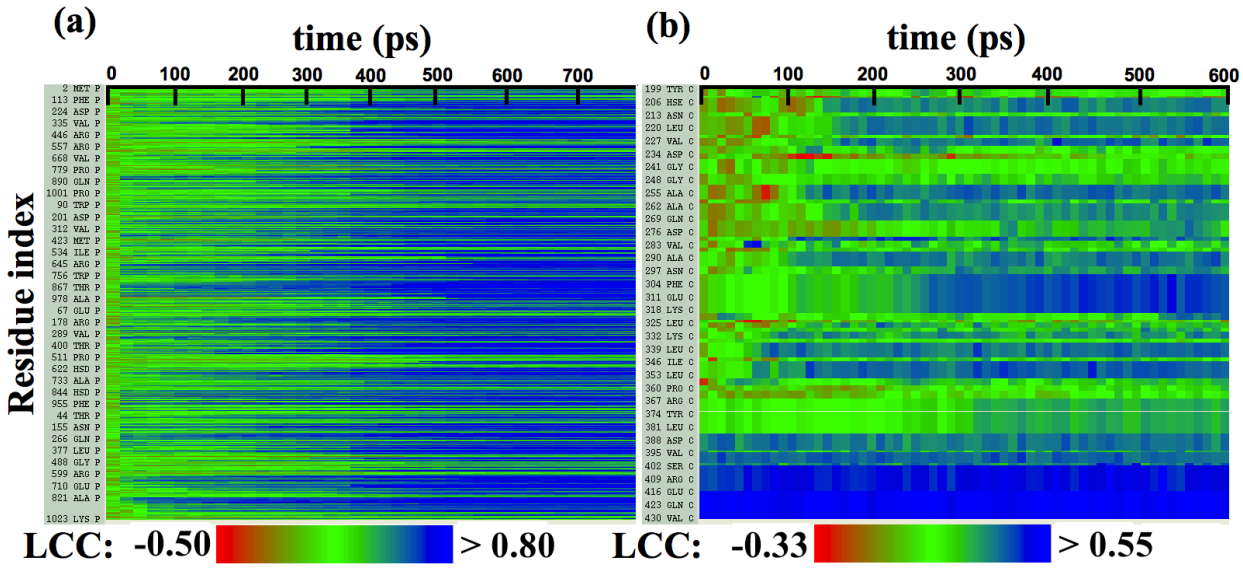

Figure 3.5: Local cross-correlations during cMDFF. Local cross-correlations of residues within the fitted regions of **(a)** β-galactosidase and **(b)** TRPV1 plotted over the course of the cMDFF fitting show improvement over the successive MDFF refinement steps.

In addition to LCC, analysis of the structures obtained through simulations included RMSD to the *de novo* structure (it is assumed that the *de novo* structures for β-galactosidase and TRPV1 are close to the "true" conformational state represented by the EM maps), EMRinger [85] scores, MolProbity [86] scores, and integrated FSCs [20].

EMRinger scores are based on whether the rotation of Cγ atoms of each residue in the

tested structure about the backbone axis produces the expected rotameric density profile in the EM map [85], and can be interpreted as a measure of both structure quality and fit. MolProbity is a measure of structure quality, rather than fit, taking into account the numbers of steric clashes, rotameric outliers, and Ramachandran outliers present in the structure [86]. It should be noted that a smaller MolProbity score indicates a better structure. Fourier shell coefficients (FSCs) are typically used to compare two maps, and is obtained, for a given radius, by calculating the correlation between the structure factors of the two maps evaluated at the given radius [20]. Typically, the degree of similarity between the two compared maps is determined by examination of the FSC profile across radii (see Section 3.4.2 for more details on FSC calculation), however one can also integrate the FSC (iFSC) between 0 Å and a pre-determined cutoff to obtain a single number indicating the degree of similarity. In the present study, iFSCs are employed as a measure of fit of a structure to a map, by constructing a simulated map of the structure, and calculating the iFSC between the simulated map and fitted map. Besides MolProbity score, other key structural metrics of a given structure can also be calculated by the MolProbity server at `http://molprobity.biochem.duke.edu/`. These metrics are also reported, but in a separate table from the fit metrics for each test case.

The purpose of employing a multitude of fitting and structure metrics is to demonstrate the robustness of cMDFF and ReMDFF to different measures of fit and structure quality. In the results presented below, cMDFF and ReMDFF are revealed to produce structures that improve over the initial ones in all the discussed measures, except for percentage of Ramachandran outliers.

3.5.3 Refinement of β-galactosidase

In the case of $\beta-galactosidase$, two initial structures were fitted to a 3.2-Å map [77] employing direct MDFF, cMDFF and ReMDFF. Since the radius of convergence of the proposed MDFF protocols was at least 7 Å for the proof-of-principle case, the initial structures were prepared with an RMSD of 7 Å from the reported structure, using the procedures described in Section 3.4.3. The first initial structure was obtained from an equilibration at room temperature. The second structure was obtained through a high temperature equilibration, and suffered from a lower local secondary structure quality. Fig. 3.2a shows the first structure superimposed on the original structure. In the Cartoon representation shown, the second structure bears a similar appearance to the first.

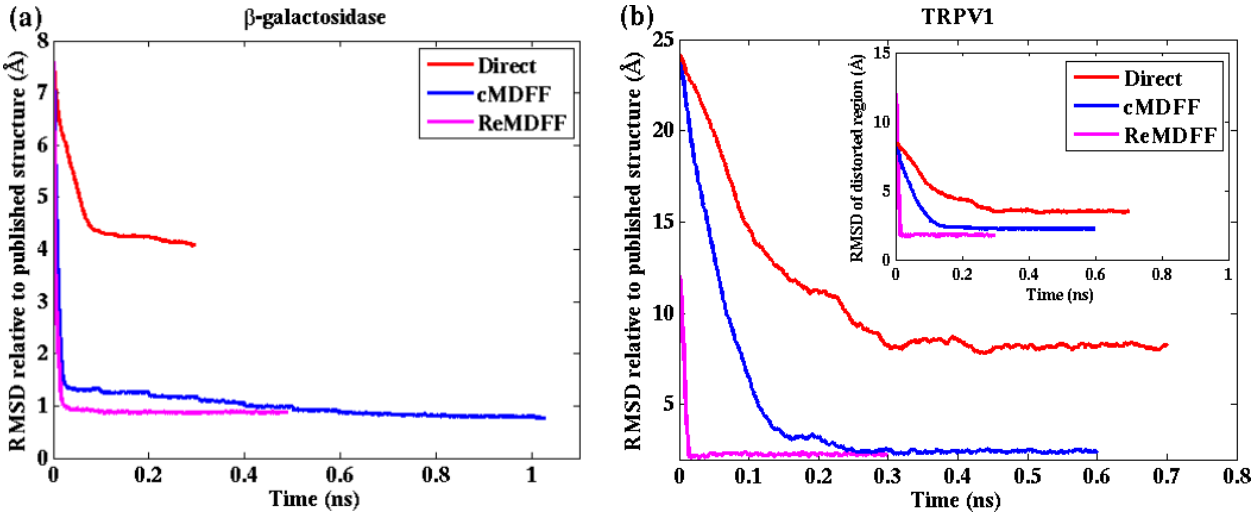

Figure 3.6: Convergence of cMDFF, ReMDFF, and direct MDFF simulations. RMSD over simulation time is plotted for the cMDFF, ReMDFF and direct MDFF simulations of **(a)** β-galactosidase and **(b)** TRPV1 monomer. RMSD is calculated with respect to the published structures (PDB 3J7H for 3.2-Å resolution and PDB 5A1A for 2.2-Å resolution). For ReMDFF, the plot contains data from a single, best-fit, replica. **(inset)** same as (b) but for the TRPV1 tetramer. For both β-galactosidase and TRPV1, cMDFF and ReMDFF outperformed direct MDFF in both efficiency and fit, as reflected in Tables 3.1 and 3.5.

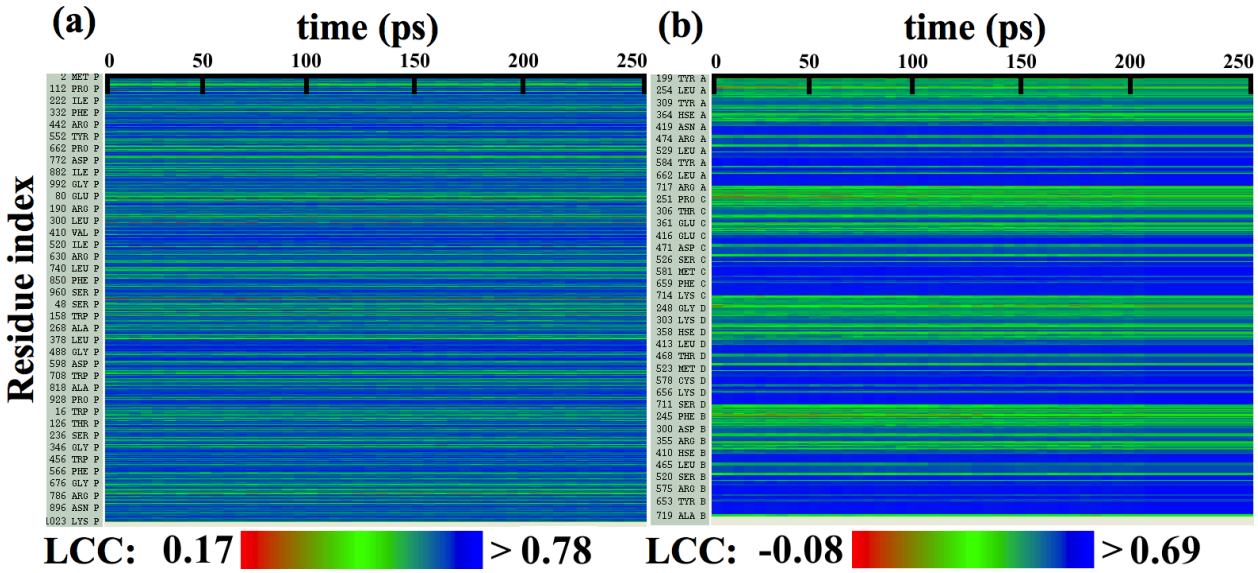

Figure 3.7: Local cross-correlations during direct MDFF to refine *de novo* structures. Local cross-correlations of residues within the fitted regions of **(a)** β-galactosidase and **(b)** TRPV1 show little change over the course of direct MDFF. The large-scale initial structures were already well-fitted within the maps. Increases in fit and structure quality of the refined *de novo* structures over the initial structures are due to local, sporadic improvements.

Fitting results for the first initial structure are summarized in Table 3.1 and show that cMDFF and ReMDFF produce final structures of higher fit and structure quality than that obtained by direct MDFF. In particular, **(i)** RMSD of the fitted structure with respect to the reported *de novo* model is 0.7 Å and 0.9 Å for cMDFF and ReMDFF respectively, much lower than the 3.7 Å RMSD attained with direct MDFF (Fig. 3.6 a); **(ii)** EMRinger scores for cMDFF and ReMDFF are 3.16 and 3.45 respectively, higher than the 1.91 obtained for direct MDFF, implying accurate fitting of sidechains into the density; **(iii)** MolProbity scores are consistently small for all the flexible fitting techniques in part due to fewer, less severe steric clashes and fewer Ramachandran outliers (further detailed in Table 3.2); **(iv)** integrated FSC (iFSC2, corresponding to the range 3.4-10 Å on the FSC plot obtained as per Section 3.4.2), considered a more stringent measure of model quality than CC [20], attained higher values of 5.22 Å and 4.66 Å for cMDFF and ReMDFF, respectively, than 2.74 Å for direct MDFF. iFSC1, evaluated at the lower resolution range of 5-10 Å improves from 2.11 Å for direct MDFF to 4.22 Å and 3.76 Å for cMDFF and ReMDFF, respectively, showing a trend similar to that of iFSC2 corresponding to the high resolution range; and **(v)** GCCs improved from an initial value of 0.48 to 0.56, 0.67 and 0.67 for direct, cMDFF, and ReMDFF protocols respectively. Similarly, typical residue LCC values improved from about 0 to greater than 0.80 (Figs. 3.5a and 3.7).

Table 3.1: **β-galactosidase MDFF results for initial structure prepared at room temperature.** cMDFF and ReMDFF provide better fitted structures than direct MDFF according to various criteria. It is noteworthy that all structures refined by any form of MDFF display an improved MolProbity [86] score compared to the original *de novo* structure.

Structure	RMSD(Å)	EMRinger	iFSC1(Å)	iFSC2(Å)	MolProb.	GCC
de novo [77]	0.0	2.25	4.03	5.00	3.14	0.67
Refined *de novo*	0.6	4.23	4.19	5.20	1.23	0.68
Initial	7.7	0.24	0.14	0.15	1.49	0.48
Direct MDFF	3.7	2.31	2.11	2.74	1.38	0.56
cMDFF	0.7	3.16	4.22	5.22	1.37	0.67
ReMDFF	0.9	3.45	3.76	4.66	1.13	0.67

Table 3.2: Structure quality indicators for β-galactosidase structures from initial structure prepared at room temperature. β-galactosidase structures investigated in the present study were uploaded to the MolProbity server (`http://molprobity.biochem.duke.edu`) to extract the quantities presented below. The results show that the cMDFF- and ReMDFF-refined structures not only exhibit good measures of fit, but also improve the clash score and rotamer geometries, relative to the *de novo* and initial structures, while incurring only a small expense in Ramachandran statistics, bad angles, and C_β deviations.

	de novo [77]	Refined *de novo*	Initial	Direct MDFF	cMDFF	ReMDFF
Clashscore	53.7	0.0	0.0	0.0	0.0	0.0
Poor rotamers (%)	11.6	3.8	4.2	3.0	4.4	1.37
Favored rotamers (%)	67.4	90.8	87.8	92.1	89.8	95.3
Ramachandran outliers (%)	0.2	0.7	2.7	3.0	1.6	2.7
Ramachandran favored (%)	97.4	95.8	91.1	91.1	94.4	90.9
MolProbity	3.14	1.23	1.49	1.38	1.37	1.13
Cβ deviations (%)	0.0	0.05	4.92	0.18	0.29	0.39
Bad bonds (%)	0.09	0.04	3.61	0.02	0.01	0.03
Bad angles (%)	0.03	0.60	3.98	0.63	0.49	0.37
RMS distance (Å)	0.007 (0.025%)	0.019 (0%)	0.035 (0.237%)	0.022 (0%)	0.019 (0%)	0.021 (0%)
RMS angle (degrees)	1.1 (0.009%)	2.2 (0.009%)	3.6 (1.177%)	2.4 (0.103%)	2.1 (0.018%)	2.3 (0.085%)
Cis prolines (%)	8.06	8.06	6.45	6.45	6.45	8.06
Cis non-prolines (%)	1.15	1.15	0.0	0.0	1.15	0.0

Overall, cMDFF and ReMDFF refinements produce structures that fit the 3.2-Å β-galactosidase map much more accurately than direct MDFF does. Fig. 3.8a visualizes the difference between the cMDFF-derived structure and the direct MDFF one in terms of fit. In judging the RMSD values to the target model the reader is reminded that equilibrium MD simulations of a single structure at room temperature typically exhibit RMSD values relative to the initial structure or the average structure of about 3 Å; the same is true for β-galactosidase. Consequently, an RMSD of 0.7 Å of the cMDFF/ReMDFF-fitted model relative to the target implies a high-quality refinement. The high quality of this refinement is further supported by visualizations of accurate sidechain placements within the density, shown in Fig. 3.9.

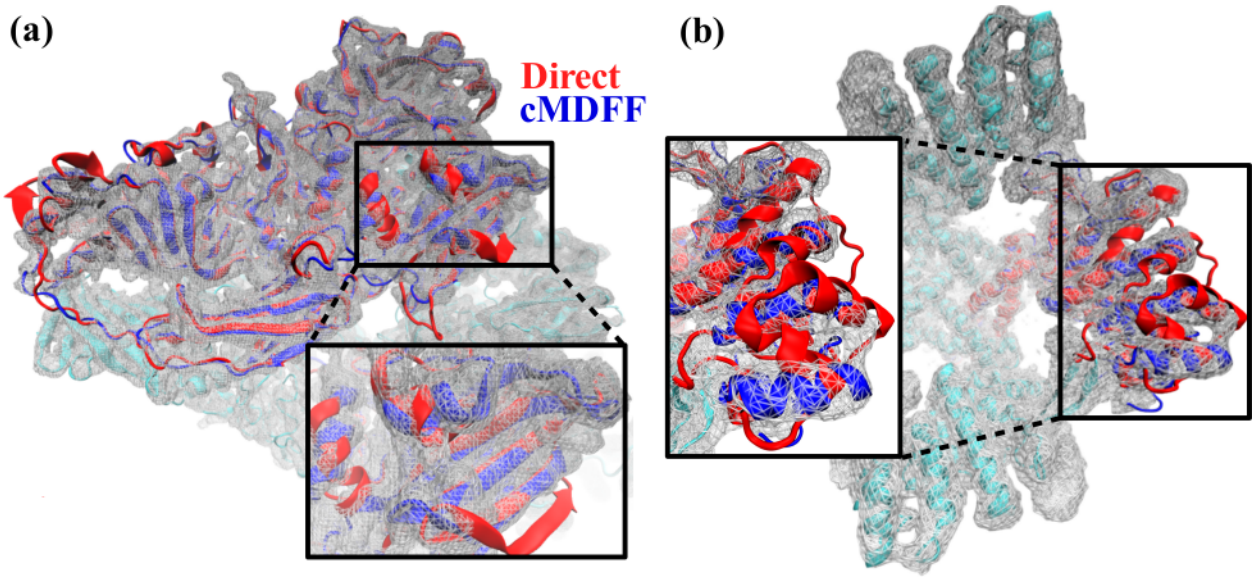

Figure 3.8: Comparison between cMDFF and direct MDFF fitted models. Models of **(a)** β-galactosidase and **(b)** TRPV1, obtained from cMDFF (blue) and direct MDFF (red) fitting simulations are superimposed. The cMDFF-fitted models fit well into the high resolution maps (grey) of each molecule, whereas the direct MDFF models have become trapped in local minima that result in portions of the models protruding from the maps. ReMDFF-fitted models are almost identical to those from cMDFF and are therefore not shown.

The cMDFF- and ReMDFF-refined structures were found to be comparable in every quality measure to the reported *de novo* structure [77]; in fact, the overall Molprobity and EM-Ringer scores are significantly better for cMDFF and ReMDFF. However, a closer look at the Molprobity score (Table 3.2) reveals that even though cMDFF vastly improves clash score and poor rotamers, it marginally increases the percentage of Ramachandran outliers and Cβ deviations relative to the *de novo* structure. Nonetheless, both cMDFF and ReMDFF improved structural statistics with respect to the initial model (Table 3.1, third row) which was intentionally chosen to have a large deviation (RMSD of 7.6 Å) from the the *de novo* structure.

In addition to simulations performed on the first initial structure, a cMDFF simulation was also performed on the *de novo* modelled structure itself, as a more realistic case comparison to the former structure. The simulation, labeled 'refined *de novo*' in Table 3.1, yielded a structure that was superior in all the quality measures considered in comparison to the *de novo* structure as well as to the structures obtained from the various MDFF fittings of the other, 7.6 Å-deviated initial model. Table 3.2 shows that not only are clash score and percentage of poor rotamers vastly improved, but the percentages of Ramachandran

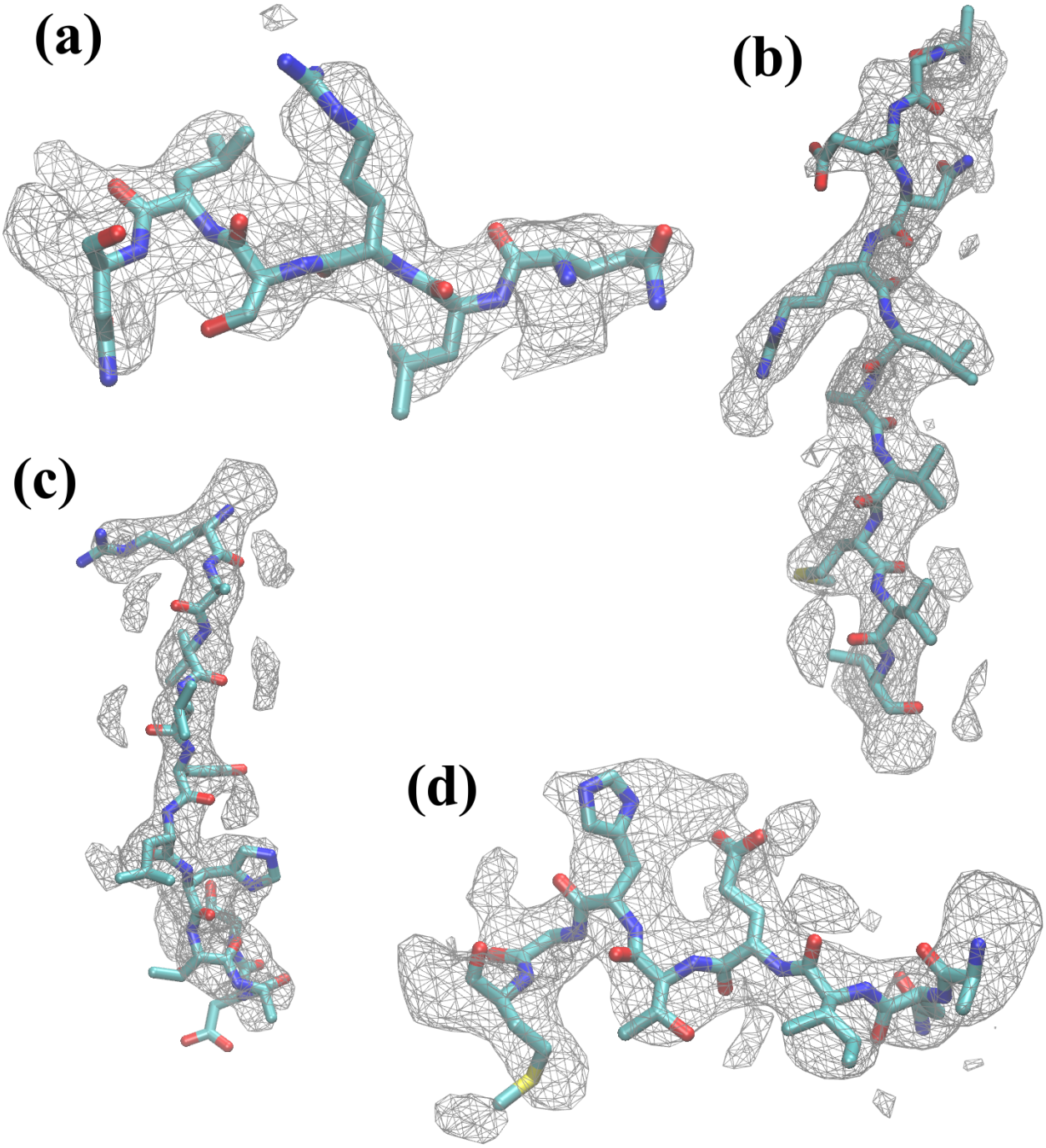

Figure 3.9: Residues of β-galactosidase fitted within density map. Several examples of residue segments, consisting of residues **(a)** 50-55, **(b)** 179-189, **(c)** 310-320, and **(d)** 413-420, are shown within the corresponding map regions. In general, both backbone and sidechains were found to have fitted well after MDFF refinement.

outliers and Cβ deviations were only slightly higher than those of the *de novo* structure itself. The small increase in the latter two measures suggest that local secondary structure quality of the obtained structure is not improved by MDFF and can be controlled instead only through the quality of the initial structure. This observation was investigated in a similar set of simulations performed on the second initial structure, which had suffered from significant secondary structure distortions introduced by the high temperature procedure used to generate it (compare Initial columns of Tables 3.2 and 3.4).

Table 3.3: Measures of fit for MDFF refinements of β-galactosidase from initial structure prepared at 1000 K.

Structure	RMSD	EMRinger	iFSC1	iFSC2	MolProb.	GCC
de novo [77]	0.0	2.25	4.03	5.00	2.33	0.67
Initial	7.6	0.26	0.10	0.09	1.16	0.25
Direct MDFF	6.2	1.91	2.11	2.73	1.21	0.47
cMDFF	3.2	2.88	3.17	3.96	1.19	0.63
ReMDFF	3.0	2.89	3.34	4.15	1.21	0.64

Table 3.4: Structural quality indicators for MDFF-refined β-galactosidase from initial structure prepared at 1000 K.

	de novo [77]	Initial	Direct MDFF	cMDFF	ReMDFF
Clashscore	53.7	0.0	0.0	0.0	0.0
Poor rotamers (%)	11.6	24.8	7.6	5.4	6.7
Favored rotamers (%)	67.4	53.7	81.4	85.0	85.4
Ramachandran outliers (%)	0.2	5.0	8.7	7.8	8.5
Ramachandran favored (%)	97.4	84.7	81.9	83.4	81.6
MolProbity	3.14	2.22	1.88	1.74	1.84
Cβ deviations (%)	0.0	20.6	0.8	0.7	3.7
Bad bonds (%)	0.09	15.4	0.03	0.02	0.02
Bad angles (%)	0.03	18.71	1.38	1.20	1.40
Cis prolines (%)	8.06	8.06	8.06	8.06	8.06
Cis non-prolines (%)	1.15	1.15	1.15	1.15	1.15

Structural statistics for the refinements of the low quality structure are provided in Ta-

ble 3.4. Relative to the initial structure, MDFF improved all measures except for the percentage of Ramachandran outliers, which has only increased despite secondary structure restraints being used. Thus, using an initial structure with a low percentage of Ramachandran outliers is crucial to maintaining the local secondary structure quality of the refined structure during any MDFF procedure. Another insight that can be gleaned from this observation is that aggregate measures of map-model validation, including RMSD (to the reported structure), GCC, and iFSC values, are insensitive to to discrepancies in local secondary structure. Despite having similar RMSD, GCC, and iFSC values, the refined *de novo* structure and MDFF refinements using the two different initial structures exhibit very different percentages of Ramachandran outliers.

In terms of efficiency, the ReMDFF protocol exhibits the quickest convergence, arriving at steady state within 0.1 ns of simulation, whereas cMDFF requires around 0.8 ns. Both methods employed eleven maps with Gaussian blurs starting from a width of 5 Å decreasing in steps of 0.5 Å towards the original reported map. To ensure that the cMDFF procedure did not over-fit the structures, cross-validation using EMRinger and FSC analysis was performed using half-maps from the EMD-5995 entry. iFSC and EMRinger values were found to be almost identical in both direct and cross comparisons. Details are provided in Section 3.4.7.

Beyond the simulations reported in Table 3.1, further simulations were performed to explore the performance of MDFF within the contexts of further types of analyses. The first of these simulations was a direct MDFF simulation of the reported β-galactosidase structure, fitting only backbone atoms to the 3.2-Å map. The fitted structure was compared to the "refined *de novo*" structure. It was found that EMRinger scores were lower (better) at 2.35 when only backbone atoms were fitted, compared to 4.23 when non-hydrogen atoms were fitted. This result suggests that even if the backbone is correctly placed, the MD force fields alone, i.e., CHARMM36 [58] here, are incapable of providing sidechain geometries consistent with the map. Refinement of the sidechains will therefore require explicit fitting to the density, above and beyond the orientations captured by the force fields alone.

In the second simulation, the resulting structure of the cMDFF simulation of the good initial structure was subjected to an equilibrium MD simulation in explicit solvent. As shown in Fig. 3.10, the equilibrium RMSD fluctuations during this simulation ranged between 3.0 Å to 3.4 Å of the starting structure. It is worth noting that these RMSD values agree well with the 3.2 Å resolution limit of the β-galactosidase map. Thus, the result indicates that uncertainties of the map resolution reflects quantitatively the structural variations of the cMDFF-fitted β-galactosidase model at the room temperature. Consequently, this model is

representative of the thermodynamic ensemble that the EM map characterizes.

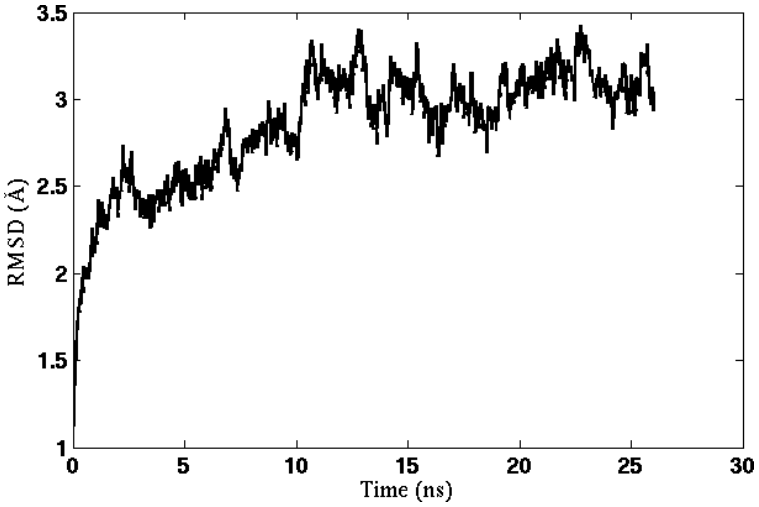

Figure 3.10: Equilibration of cMDFF-refined model of β-galactosidase. The model resulting from a cMDFF fitting of β-galactosidase to the 3.2-Å map is subject to an equilibration MD simulation. The RMSD plot of the structure shows that it converges within 10 ns to an RMSD value of 3 Å.

The third set of simulations takes advantage of a unique opportunity, presented by the availability of two different maps of the same structure, at resolutions of 3.2 Å and 2.2 Å, to compare the results of fitting β-galactosidase to maps of different resolutions. The reported structures were subjected to direct MDFF simulation for 0.7 and 1 ns for the 3.2-Å and 2.2-Å models, respectively. The RMSF for each residue is calculated over consecutive 10-ps windows during the fitting. The RMSF values for all residues, including those for the PETG binding pockets [78], are plotted in Fig. 3.11, reflecting smaller fluctuations during the fitting to the 2.2-Å map than in the 3.2-Å one. The relationship between fluctuation and map quality is examined in greater detail in Results, and imply that the RMSF of the fitted structure correlates negatively with the resolution of the corresponding map.

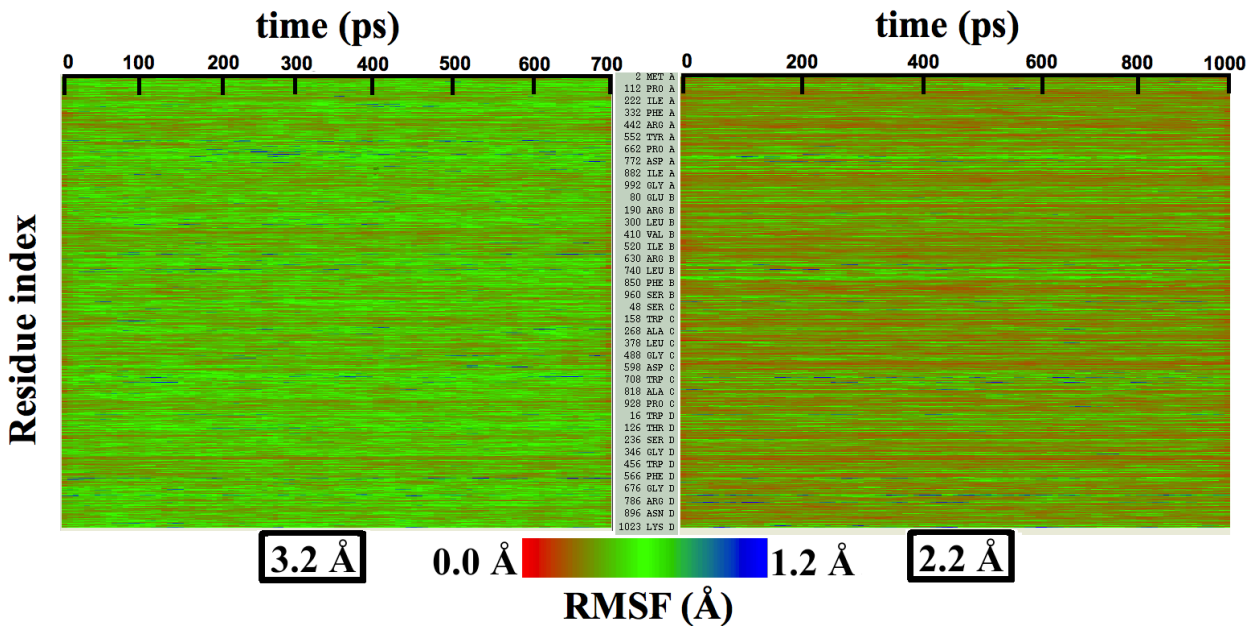

Figure 3.11: RMSF values of individual residues during direct MDFF of published β-galactosidase models. Published models corresponding to the 2.2-Å and 3.2-Å maps of β-galactosidase are fitted to their respective maps using direct MDFF. The RMSF values of all the residues along the protein sequence are plotted showing those from the 2.2-Å map are lesser than those from 3.2-Å map.

3.5.4 Refinement of TRPV1

The initial structure for TRPV1 was fitted using direct MDFF, cMDFF, and ReMDFF protocols to the reported 3.4-Å map [76]. To obtain this initial structure, an interactive MD procedure was used to distort one subunit of the *de novo* structure, as described in Section 3.4.3, such that the total deviation of the initial structure from the *de novo* structure was 10 Å in RMSD, and that of the distorted subunit was 25 Å. The latter degree of deviation is in the ballpark of the lowest resolutions of usable EM maps and, therefore, represents the upper limit of uncertainty between an initial structure and the fitted structure that MDFF can still reconcile. Thus the present example represents an extreme test case for evaluating the radius of convergence of the proposed MDFF methods.

Table 3.5: TRPV1 MDFF Results. Similarly to the simulations of β-galactosidase (Table 3.1), cMDFF and ReMDFF refinement of TRPV1 produce the best fitted models as analysed by various metrics. The models obtained through both cMDFF and ReMDFF are better or equal to the *de novo* structure in every analysis. Numbers in parentheses are representative of the whole structure (tetramer) while the main entries refer only to the single monomer that was fit.

Structure	RMSD	EMRinger	iFSC1	iFSC2	MolProb.	GCC
de novo	0.0	0.83	2.33	2.62	3.80	0.53
			(3.48)	(4.33)		(0.72)
Refined *de novo*	1.1	1.75	1.90	2.13	1.52	0.54
			(3.52)	(4.42)		(0.73)
Initial	24.1	0.62	0.67	0.77	3.79	0.16
			(2.96)	(3.81)		(0.67)
Direct MDFF	7.9	1.51	1.61	1.79	1.71	0.50
			(3.68)	(4.47)		(0.72)
cMDFF	2.4	1.68	2.37	2.62	1.60	0.54
			(3.66)	(4.43)		(0.72)
ReMDFF	2.5	1.99	2.41	2.75	1.47	0.53
			(3.68)	(4.50)		(0.73)

Table 3.6: Structure quality indicators for TRPV1 structures. TRPV1 structures investigated in the present study were uploaded to the MolProbity server (`http://molprobity.biochem.duke.edu`) to extract the quantities presented below. As in the case of β-galactosidase, the overall MolProbity score has been improved, relative to the *de novo* and initial structures, by cMDFF and ReMDFF at the expense of a small increase in Ramachandran outliers and Cβ deviations.

	de novo [76]	Refined *de novo*	Initial	Direct MDFF	cMDFF	ReMDFF
Clashscore	92.8	0.0	91.6	0.0	0.0	0.0
Poor rotamers (%)	28.8	3.1	28.8	2.4	2.8	4.4
Favored rotamers (%)	53.8	90.5	53.8	92.8	91.4	87.6
Ramachandran outliers (%)	1.0	3.4	1.0	3.5	3.5	3.4
Ramachandran favored (%)	94.5	92.3	94.5	90.8	91.1	92.3
MolProbity	3.92	1.34	3.91	1.32	1.36	1.47
Cβ deviations (%)	0.0	0.53	0.0	0.26	0.32	1.01
Bad bonds (%)	0.72	0.0	0.77	0.0	0.0	0.0
Bad angles (%)	0.52	0.42	0.51	0.41	0.37	0.50
RMS distance (Å)	0.019 (0%)	0.019 (0%)	0.017 (0%)	0.019 (0%)	0.019 (0%)	0.019 (0%)
RMS angle (degrees)	1.9 (0.018)	1.9 (0.018%)	2.0 (0.240%)	1.9 (0.014%)	1.9 (0.014%)	1.9 (0.032%)
Cis prolines (%)	15.38	15.38	15.38	15.38	15.38	15.38
Cis non-prolines (%)	0.62	0.62	0.62	0.62	0.62	0.62

Fitting results for TRPV1, as shown in Table 3.5, were significantly better for cMDFF and ReMDFF than for direct MDFF. Fig. 3.8b illustrates the contrast in fit between the cMDFF and direct MDFF-derived structures. In addition to the simulations applied to the distorted initial structure, a cMDFF simulation of the *de novo* structure, labelled as 'refined *de novo*', was also performed to represent a realistic scenario where starting structures typically do not stray as far from the map as the present distorted structure does. In summary, **(i)** RMSD of the fitted structure with respect to the reported *de novo* structure is 7.9 Å for direct MDFF, higher than the 2.4 Å and 2.5 Å RMSD values for cMDFF and ReMDFF, respectively (Fig. 3.6 b); **(ii)** EMRinger scores for cMDFF and ReMDFF are 1.68 and 1.99 respectively, higher than the 1.51 score obtained for direct MDFF; **(iii)** MolProbity scores [86] are 2.4 and 2.5 for cMDFF and ReMDFF, smaller than the 7.9 score for direct MDFF, implying

fewer, less severe steric clashes and fewer poor rotamers in the former than in the latter; **(iv)** integrated FSC (iFSC2, for the range 3.4-10 Å obtained as described in Section 3.4.2), attains higher values of 2.62 and 2.75 for cMDFF and ReMDFF respectively, than the 1.79 value for direct MDFF. iFSC1, corresponding to the lower resolution range of 5-10 Å was found to behave similarly to iFSC2; and **(v)** GCCs improved from an initial value of 0.16 to 0.50, 0.54 and 0.53 for direct, cMDFF, and ReMDFF protocols, respectively. Similarly, typical residue LCC values improve from 0 to 0.5 or higher, as shown in Fig. 3.5 b.

Measures of structural quality for the above fits are tabulated in Supplementary file 3.6. As the table shows, the percentage of Ramachandran outliers remained low across all simulations for TRPV1, primarily because the initial structure did not suffer from significant local secondary structure defects. Cross-validation with half-maps was also performed on the cMDFF structure to ensure that it was not over-fitted. As in the case of β-galactosidase, iFSC and EMRinger scores for direct and cross comparisons were similar. FSC analysis results are described in Section 3.4.7.

Like in the case of β-galactosidase, the ReMDFF protocol exhibited the quickest convergence, arriving at steady state within 0.02 ns of simulation, whereas cMDFF required around 0.27 ns. It is worth repeating at this point that the TRPV1 set of simulations demonstrate the large radius of convergence of 25 Å of cMDFF and ReMDFF.

A separate set of model validation analyses was performed on the well-resolved transmembrane (TM) portion (residues 381 to 695) of TRPV1 to allow direct comparison of a MDFF-refined model with one from Rosetta [85]. The TM region had previously been refined employing Rosetta tools [103], providing an opportunity for comparison. Two MDFF simulations were performed, the first with only non-hydrogen sidechain atoms coupled to the density and harmonic restraints holding backbone atoms in the configuration of the reported structure [76], and another with all non-hydrogen atoms coupled to the density and the backbone restraints removed.

MDFF characteristics for fitting the isolated TM region of TRPV1 are summarized in Table 3.7. Quality of fit measures, namely EMRinger and iFSC, for the backbone-restrained simulations were lower than those of the Rosetta-derived structure. However, MolProbity scores for the MDFF-derived structures are better than those of Rosetta. Allowing the backbone to be fitted into the map without restraints from the reported structure substantially improved the quality of fit measures so that they are comparable to those of Rosetta's, while maintaining a lower MolProbity score.

Table 3.7: MDFF for the TRPV1 TM region. Three MDFF refinements of the TRPV1 TM region were performed under different conditions. Measures of quality of the resulting structures are compared with those for the published structure and for the structure obtained from the Rosetta software. The tabulated results show that MDFF with backbone atoms free to move fared as well or better than Rosetta in terms of the measures considered.

Structure	RMSD	EMRinger	iFSC1	iFSC2	MolProb.	GCC
de novo	0.0	1.05	3.20	4.28	4.02	0.63
Rosetta	1.2	2.57	3.55	4.60	1.55	0.63
Backbone restrained	1.8	2.34	3.59	4.71	1.08	0.63
Backbone free	1.2	2.51	3.80	4.85	1.37	0.64

3.5.5 MDFF Protocol Efficiency

The rate of convergence to a final structure for cMDFF and ReMDFF are compared against direct MDFF for the TRPV1 and β-galactosidase cases. Fig. 3.6 shows the time evolution of RMSD relative to the *de novo* structure for cMDFF, ReMDFF, and direct MDFF for the two proteins. It should be noted that the plots do not include the final refinement step, which is the same across all three protocols.

cMDFF and ReMDFF reach similar RMSD levels, outperforming direct MDFF. ReMDFF converges more quickly than cMDFF in both examples. Of the 6 replicas employed for the ReMDFF of TRPV1, two resulted in poorly fitted structures, having become trapped in density minima even after exchanging with the lowest-resolution map, i.e. $\sigma = 5$ Å. All replicas are monitored during simulation and poorly fitted ones can be discarded by a user.

It is also worth noting that the region of the TRPV1 map to which both cMDFF and ReMDFF successfully fitted the structure (residues 199 to 430) is characterized by a diverse range of local resolutions from 4 Å to 6 Å and poses a challenge to the conformational sampling capability of any flexible fitting technique. For the same reason, this region was avoided during Rosetta refinements of TRPV1 [20], but is addressed now via MDFF.

3.6 RMSF Analysis

An EM density map represents a thermodynamic ensemble of atomic conformations [14, 104, 105]. Conventionally, however, only a single structure representing a best fit to the map is reported, begging the question of how statistically representative a single model can

be. To quantify the deviation of a fitted structure from the rest of a simulated ensemble of molecules, root mean square fluctuation (RMSF) of the structure relative to the ensemble-averaged structure was computed during an MDFF refinement simulation.

In this section, the RMSF of a structure being fitted is first shown to be complementary to other metrics in evaluating the quality of fit of the model, as well as to represent the degree of natural conformational variation within the thermodynamic ensemble underlying the map. Second, the RMSF values are found to correlate both locally and globally with the resolution of an EM map, providing an interpretation of map quality based on the inherent (i.e., natural) dynamics of the macromolecule under observation. Finally, RMSF values are also employed to identify optimal B-factor values for the sharpening of a map. Altogether, the results of the present study demonstrate that RMSF of a fitted structure during an MDFF refinement provides valuable information on the structure.

RMSF analysis is performed as follows. The local resolutions of a density map can be computed with ResMap [106] and used within VMD to select the atoms of a structure that are contained in a range of resolutions found by the ResMap analysis. The average local RMSF of each selection can then be calculated over an MDFF simulation, after the structure has stabilized. In principle any criteria for atom selection can be used for RMSF analysis, though we use local resolution of the EM density here to illustrate the correlation between the two measurements. Additionally, we compute a global average RMSF of the entire structure.

The ensemble-based nature of the RMSF analysis means that the quality metric is not dependent on a single structure, but instead a large family of structures can be employed as a better representative of the data. Ensemble-based analyses are a natural and powerful benefit of the MD-based nature of MDFF. RMSF analysis does not, however, require MDFF to be used as the method of refinement. In principle, any refinement method can be used to obtain the fitted model, before a subsequent short MDFF simulation of the fitted model is performed to obtain the data necessary for the RMSF analysis. For the present study, the output consists of an average RMSF value over each atom selection in the structure representing a given range of local resolutions in the corresponding EM map. Fig. 3.12 shows the correspondence between average RMSF values and local resolutions in the β-galactosidase and TRPV1 structures.

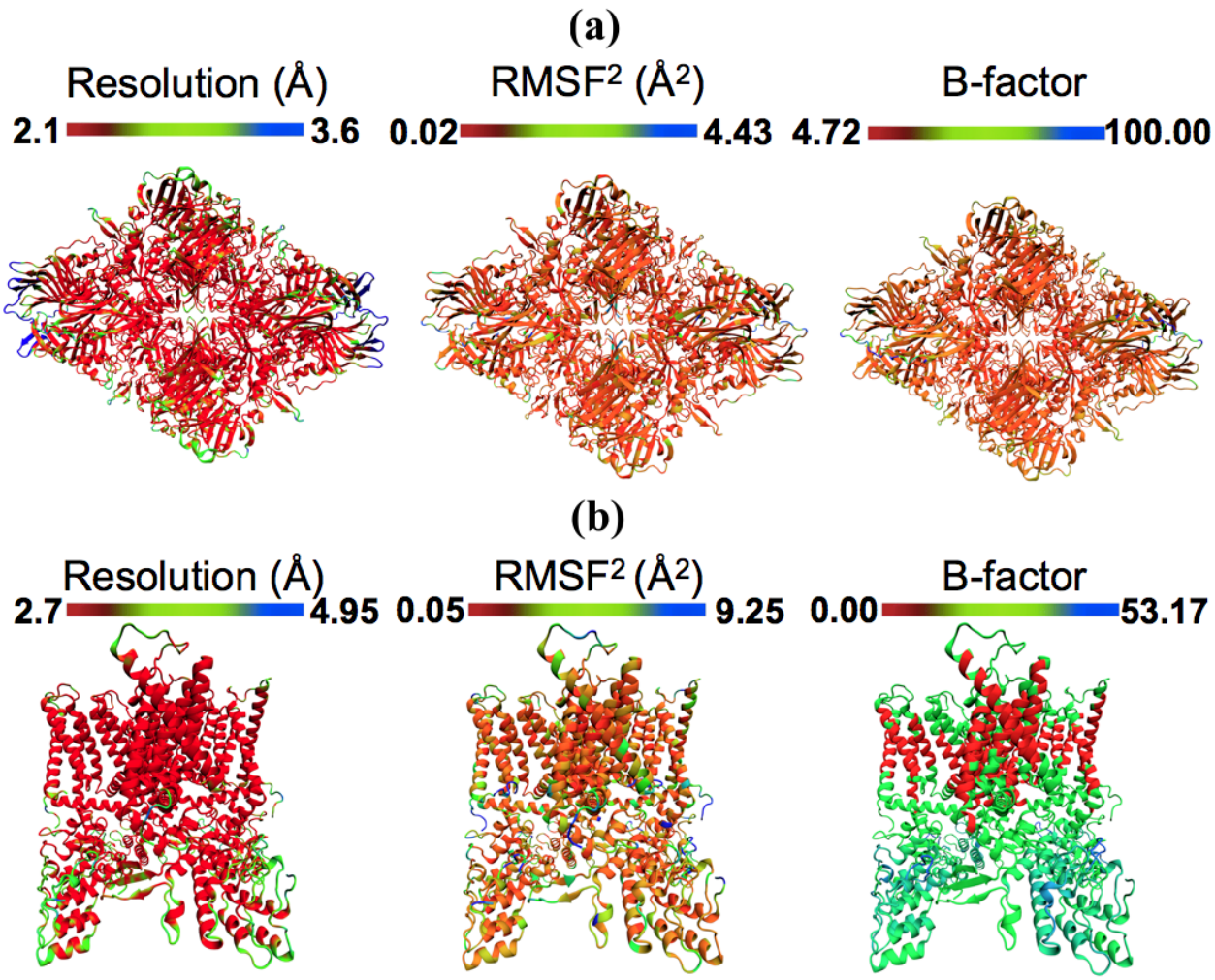

Figure 3.12: Models colored by local resolution, square of RMSF, and B-factor. The published structures of **(a)** β-galactosidase (PDB 5A1A) and **(b)** TRPV1 (PDB 3J5P) are colored by the local EM map resolutions, the per-residue mean square fluctuations (RMSF^2) during MDFF simulation, and published B-factors. Comparison of these figures show qualitative agreement between local resolution, RMSF^2, and B-factor. In fact, the local resolutions and B-factors correlate linearly with RMSF^2 of a fitted model both in the presence as well as absence of the EM map.

3.6.1 RMSF and Quality of Fit

RMSF analysis was applied to the cMDFF refinement of β-galactosidase. The RMSF of each residue was tracked over the course of the simulation and shown in the color plot in Fig. 3.13. The high RMSF values at the beginning of the simulation reflects a diverse ensemble of poorly fit structures, exemplified by the initial structure (see "Initial" row in Table 3.1). The structure explores this ensemble in the early phase of the fitting, before

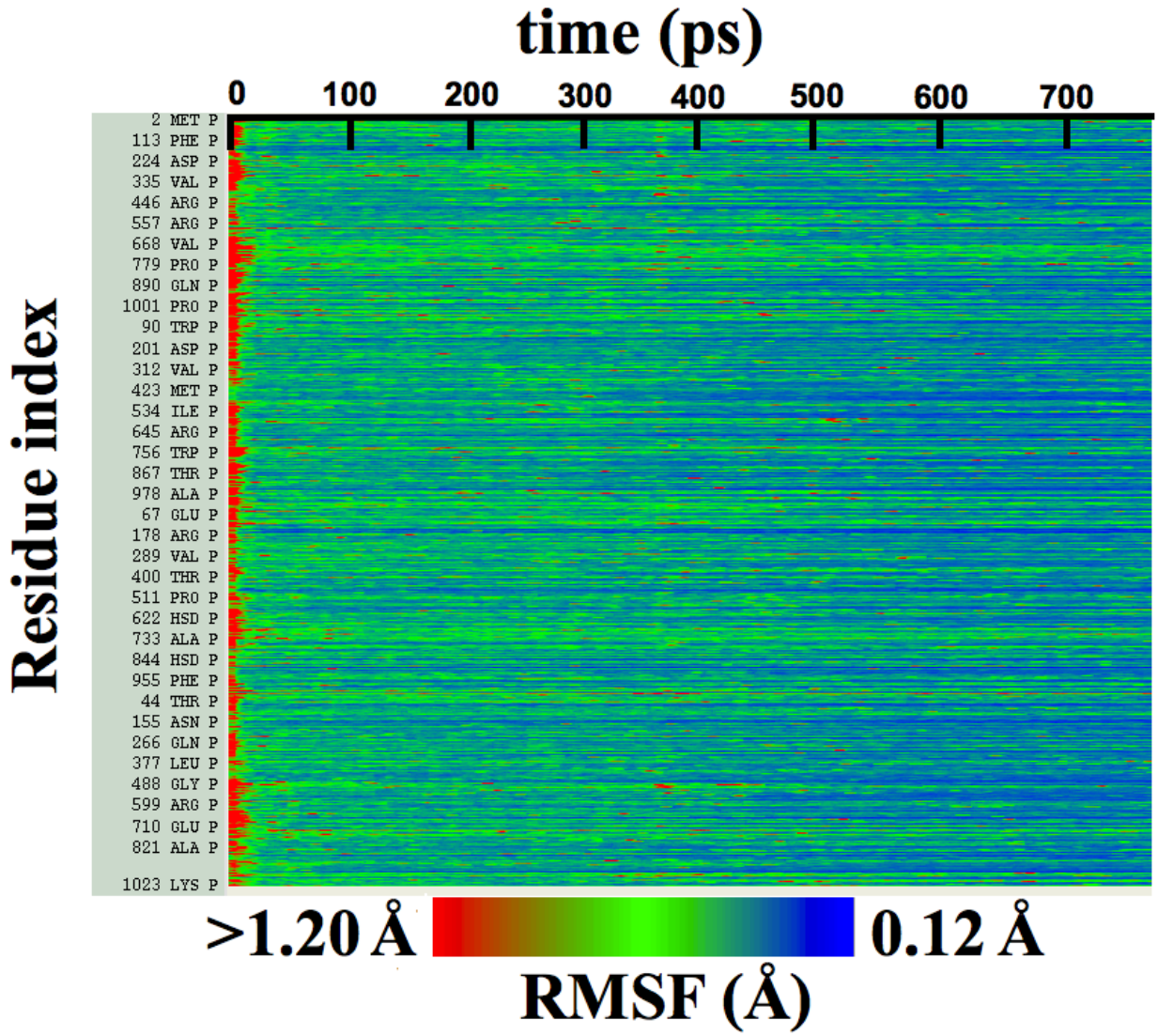

Figure 3.13: Per-residue RMSFs over β-galactosidase cMDFF fitting. Residue RMSFs as a function of progress of cMDFF fitting show a general trend of decrease as the structure becomes better fit.

converging to a smaller ensemble, characterized by low RMSF values, within the confines of the map as the simulation progresses.

The initial conformation is a poor fit of the map, characterized by low values of GCC, LCC and iFSC (the row containing 'initial' structure in Table 3.1). Such conformations belong to a diverse ensemble of poorly fit structures, explored by the structure in the early phase of the fitting, that gives rise to high initial RMSF values shown in Fig. 3.13.

The results suggest that low RMSF values indicate **(i)** the structure has been modelled

unambiguously within the map, and **(ii)** the structure can be regarded as representative of the ensemble underlying the 3.2-Å β-galactosidase map. Conversely, high local RMSF values would indicate that the residues are outside the map potential and hence poorly fitted, or that the local map region has a low resolution.

3.6.2 RMSF and Quality of Map

Apart from representing the quality of fit, RMSF values during an MDFF simulation correlate closely with the overall and local resolution of an EM map. Even for high-resolution cryo-EM data, resolution is not always uniform throughout a map. For example, Fig. 3.12 shows the variation in local resolution of map regions, coded as colors of the residues in the β-galactosidase and TRPV1 structures. Conformational flexibility can cause heterogeneity in the cryo-EM data [107], producing local resolutions lower than that of the overall map.

Local resolution analysis [106] can be especially important for determining the parts of a high-resolution map that realistically contain side chain information and the parts that do not, preventing over-interpretation of the latter. MDFF protocols can be adjusted to account for such local variations and better inform the process of model validation. For example, where side chains are not fully resolved, the side chain atoms can be decoupled from the MDFF potential during fitting. One may also weight the contributions of atoms in low-resolution regions less than in high-resolution regions when calculating the overall cross-correlation. These suggestions have not been implemented in the present study, but even as a simple tool by itself, local resolution analysis can provide a gauge of the spatial uncertainty in different regions of a fitted structure by virtue of the local map resolution.

To demonstrate the RMSF-resolution correlation, MDFF simulations of several test molecules were performed, including TRPV1 (PDB 3J5P) , β-galactosidase modelled from the 2.2-Å map (PDB 5A1A), γ-secretase (PDB 5A63 and 4UPC), and the T20S Proteasome (PDB 3J9I). It was found that the lower the overall resolution of the map, the higher the corresponding overall RMSF during MDFF simulations. For example, the overall RMSF during MDFF of the 4.5-Å γ-secretase model and map is greater than that of the 3.4-Å model and map which, in turn, is greater than that of the 2.2-Å model and map of β-galactosidase (see RMSF labels on the upper column of Fig. 3.14). The correlation between map resolution and model RMSF extends to local features within the density. In Fig. 3.14 (upper row) and Fig. 3.15 (for the proteasome case), RMSFs of atoms are linearly correlated with local resolutions of the corresponding map regions.

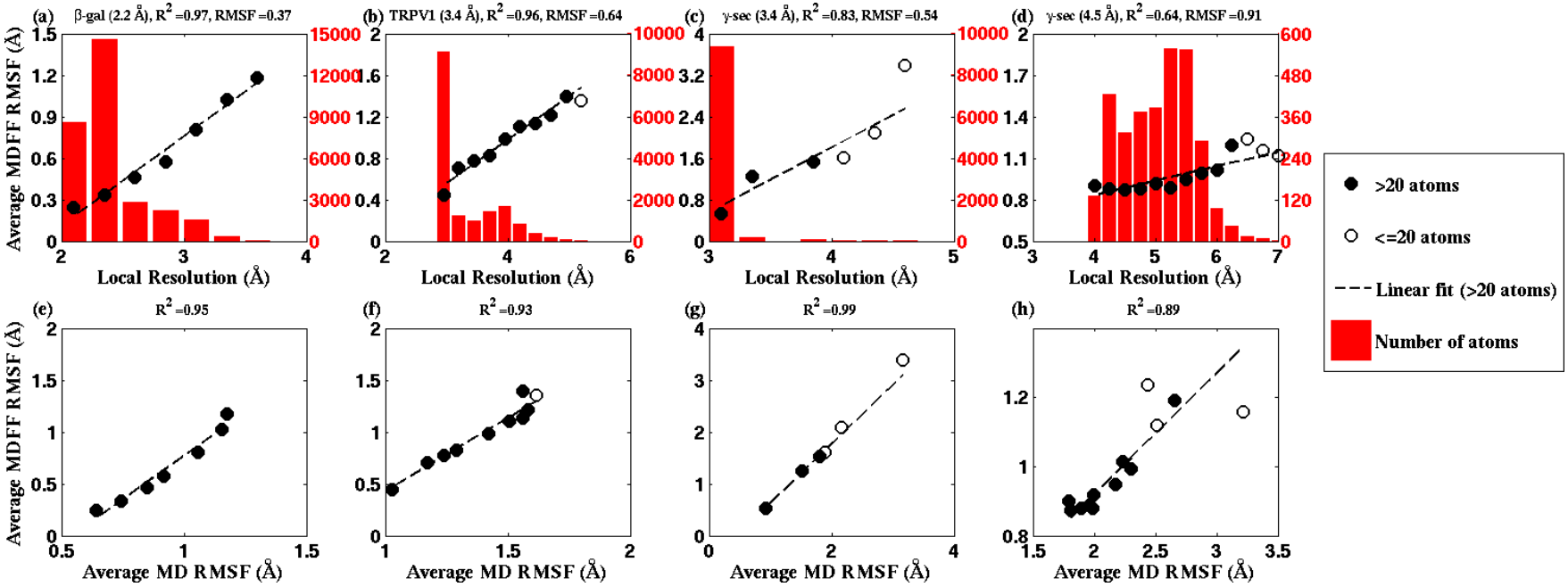

Figure 3.14: RMSF vs. local resolution plots for various simulations. For each test case shown, atoms in the MDFF-refined structure are grouped by local resolution of the map regions they are fitted into. The average RMSF value of atoms (during MDFF simulation) in each resolution bin is calculated and plotted against the local resolution in the cases of **(a)** β-galactosidase (β-gal) at 2.2 Å,**(b)** TRPV1 at 3.4 Å, γ-secretase (γ-sec) at **(c)** 3.4 Å and **(d)** 4.5 Å resolution, and proteasome (see Fig. 3.15). The numbers of atoms in the resolution bins are displayed as a histogram (in red) spanning a system-specific range of resolutions. The lowest resolution bins contained low (< 20) populations and visual inspection consistently revealed the atoms to be on the edges of the density, and were therefore ignored during further analysis. A clear linear correlation between RMSF and local resolution can be found in each case. Applying a linear fit produces the high R^2 value shown in each graph heading. Also displayed in each heading is an overall RMSF, averaged over all atoms in the system. The overall RMSF reflects the conformational variety of structures that fit within the map, and is found to correspond to the map resolution such that higher resolutions produce lower RMSFs. The second row of plots show that the RMSF during MDFF simulation also linearly correlates with RMSF during unbiased MD simulations of **(e)** β-galactosidase, **(f)** TRPV1 and **(g,h)** γ-secretase, establishing that fluctuations during MDFF reflect the inherent flexibility of a system.

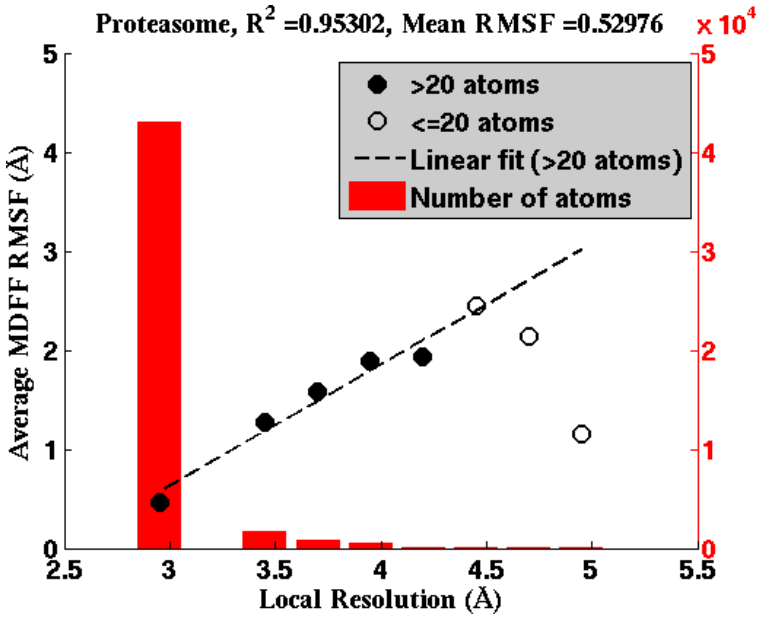

Figure 3.15: Average RMSF vs. local resolution during MDFF simulation of proteasome. In the proteasome test case, the average RMSF of atoms corresponding to each local resolution, determined by ResMap, correlates linearly with the resolution. The same correlation was observed in all test cases considered (see Results).

The linear correlation between RMSF and local resolution persists even for unbiased MD simulations (Fig. 3.14, bottom row). In the absence of the MDFF potential, the local RMSF value can be attributed to the flexibility of the corresponding region on the structure [42]. Taken together with the RMSF-resolution correlation in the MDFF simulations, this observation suggests that flexibility of the molecule during the imaging process is a key contributor to the resolution of the resulting image.

In summary of the results presented so far, the present study establishes that RMSF, together with GCC, LCC, EMRinger [85], and iFSC, provide a comprehensive set of criteria for evaluating model and map quality on both global and local levels. The added value of RMSF is particularly evident on the local level, where the other measures may not perform as consistently. For example, a high LCC may be the result of a highly flexible structure fitting to a low-resolution region of the map, and not necessarily of a good representation of the local structure. As a result, although multiple low-resolution regions of the model in Fig. 3.16 a possess similar LCCs, disparate RMSFs of the same regions clearly indicate differences in local quality of the model. Likewise, EMRinger scoring, when applied to small groups of residues, does not correlate with local resolution (Fig. 3.16 b), and, therefore, is incapable of distinguishing between regions of small numbers of atoms of varying local model

quality. In contrast, RMSF clearly resolves the local resolution and thus, resolves the map and model quality of regions even with as few as 100 atoms.

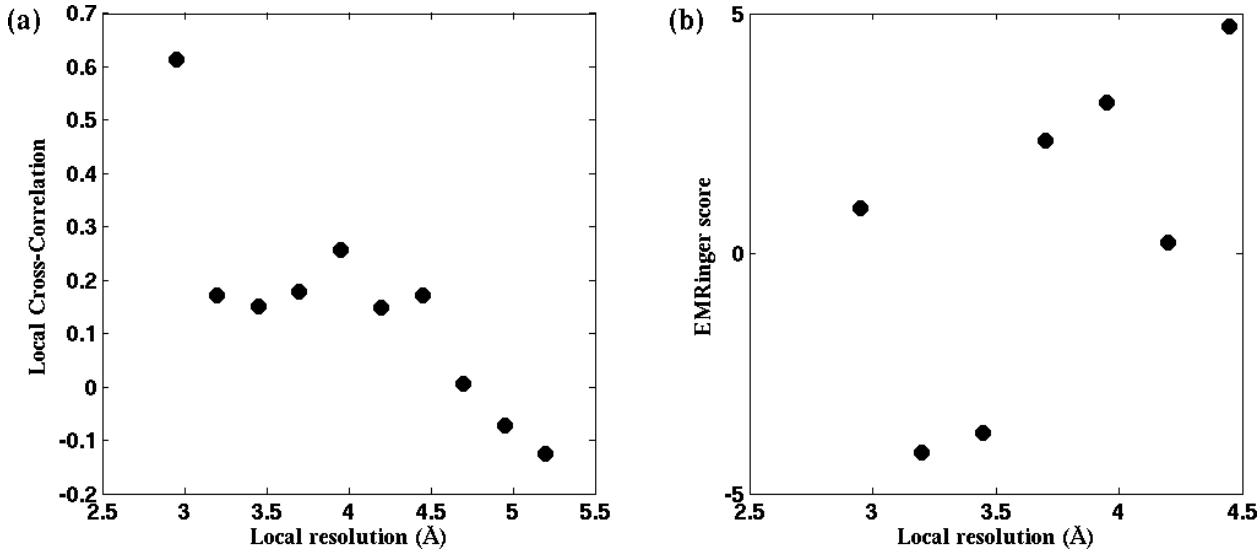

Figure 3.16: EMRinger score and LCC do not predict local resolution in TRPV1. **(a)** Local cross-correlation and **(b)** EMRinger scores obtained from residues of a fitted model of TRPV1 do not exhibit one-to-one correspondence to local map resolutions.

3.7 Conclusion

The new MDFF variants, cMDFF and ReMDFF, have been shown to ameliorate direct MDFF's shortcomings in cases where a high-resolution (sub-5 Å) map is used for fitting. In addition, cMDFF and ReMDFF also increase the radius of convergence to at least 25 Å, which is at least twice that reported for Rosetta refinements of the 20S proteasome [20]. These capabilities are exemplified in the cases of β-galactosidase and TRPV1 reported in this chapter.

ReMDFF is related to cMDFF through its use of multiple maps of different resolutions. However, the replica-exchange mechanism used by ReMDFF enables a higher degree of automation than in cMDFF - users are not required to track the progress at each step of the fitting, allowing fast and hands-off fitting. Furthermore, the ReMDFF algorithm is amenable to parallel implementation, particularly on a cloud computing platform such as Amazon Web Services [108], and is thus accessible to researchers who do not have adequate computing resources or access to their own infrastructure.

In the course of the present study, fitted structures were evaluated using multiple metrics that reflect both the quality of fit and quality of the structure. Furthermore, the argument was made for including in structure evaluation locally evaluated metrics, e.g. LCC, since global metrics may fail to identify local fitting errors, which occur more frequently in the case of high-resolution maps.

Finally, the present study proposes RMSF as an addition to the list of metrics for evaluating fitted structures. RMSF offers several advantages over the other metrics, including recognition local segments in the structure that contain more uncertainty in position, distinction between map quality and fit quality as a source of uncertainty, and robustness as a result of being a measure of an ensemble rather than a single structure.

While there are few high-resolution cryo-EM maps reported to date, these maps are expected to become more commonplace as cryo-EM and even other technologies such as X-ray free-electron lasers [109] mature. The proposed methods and techniques in this study will pave the way for the combination of MD and high-resolution maps to extract finely detailed information on macromolecular structure and dynamics.

CHAPTER 4

ADAPTIVE MULTILEVEL SPLITTING IN SIMULATIONS OF DRUG DISSOCIATION[1]

Adaptive Multilevel Summation (AMS) is a rare event sampling method that requires minimal parameter tuning and that allows unbiased sampling of transition pathways of a given rare event. Here, AMS is applied to molecular dynamics (MD) simulations to measure the rates of events that occur on far longer time scales than traditional MD is able to access. This chapter describes the implementation of AMS in NAMD, validation of the algorithm using a simple idealized system and subsequently an actual biological test case, and finally the insights drawn from the study that are instructive for future development.

4.1 Introduction

The difficulty of simulating systems over long time scales is arguably the greatest limitation of MD. Fortunately, this limitation can be overcome in many cases with the use of advanced sampling techniques. For example, free energy methods, such as umbrella sampling [25], free energy perturbation [110, 111, 112, 27], metadynamics [113, 26], and adaptive biasing force [114, 28], and variants thereof [115, 116, 117, 118, 119, 120], apply external forces on the system to expedite exploration of the regions of interest in the free energy landscape of the system. However, the external forces applied to the system may result in states that are not representative of the dynamics of the processes being studied (see Kopelevich 2013 [121] for an example).

Another class of techniques directly sample reaction paths, rather than the distribution of intermediate states. These techniques involve the setting up of branching points along a chosen reaction coordinate, at which trajectories are initialized. At each branching point, the sampling method builds upon the reference prior probability of the branching point and

[1]The research presented in this chapter has been published in D. Aristoff, T. Lelièvre, C. G. Mayne, and I. Teo, *ESAIM Proc. Surv.*, **48** (2015), p. 215–225, and I. Teo, C. G. Mayne, K. Schulten, and T. Lelièvre, *J. Chem. Theor. Comp.*, **12**(6) (2016), p. 2983–2989.

focuses on sampling the posterior probabilities of trajectories emanating from the branching point to the next branching point, thus avoiding the difficulty of sampling a very small overall unconditional probability of the event. Members of this class of techniques include transition interface sampling [122, 30], forward flux sampling [123, 31], and multilevel splitting [29].

Adaptive Multilevel Splitting (AMS) [23, 24] is a variant of multilevel splitting, designed to minimize the need for prior knowledge, such as good choices of reference probabilities in importance sampling or branching locations and frequencies in transition interface sampling, forward flux sampling and multilevel splitting, by adaptively determining the branching points during the simulation. In contrast to most other rare event sampling methods, AMS does not require prior definition of branching point locations and frequencies, thus enabling easy implementation to a potentially high degree of automation even for processes that are highly complex and/or for which little information apart from the initial and final states is available.

This chapter presents a proof-of-principle of the applicability of AMS to MD simulations. The first set of validation simulations measured the rate of escape of a single particle from a potential well. The measured rate was in excellent agreement with both direct measurement using traditional MD and calculations using solutions of the Smoluchowski equation. Subsequently, AMS was applied to an actual biological system - the benzamidine-trypsin complex, where the AMS-measured rate of dissociation of benzamidine from trypsin was found to agree with the experimentally-determined rate.

The benzamidine-trypsin study is motivated by the wider context of drug efficacy prediction *in silico*. In particular, the drug residence time is increasingly being seen as a major determinant of potency [124, 125]. As such, there have been efforts to develop different techniques to measure drug residence times, or equivalently, the dissociation rate, through MD simulations [126, 127, 128, 129].

The complexity of unbinding processes underlies the difficulty of obtaining dissociation rates. In the case of benzamidine-trypsin, previous computational studies [127, 128, 129] have identified multiple dissociation pathways and utilized Markov State Models [130, 131] to characterize the entire dissociation process and obtain estimates of the overall dissociation rate. In the present study, AMS is applied along a simple reaction coordinate to estimate the dissociation rate. In contrast to the other computational studies, prior determination of pathways and specific metastable states was not required in the AMS calculation, but it is noted that such knowledge may be helpful in obtaining better convergence of results. It should also be noted that the retrospective reconstruction of pathways and metastabilities

is possible through the reactive pathways obtained in the AMS algorithm [24], but has not been undertaken in the present study.

4.2 Adaptive Multilevel Splitting

This section describes the basic AMS formulation and the implementation of the algorithm in NAMD.

4.2.1 Basic Algorithm

The output of AMS is the committor probability, defined as the probability that a system, after leaving a given initial state, reaches the given final state before returning to the initial state. Following the formulation by Cérou and Guyader [23], let $\{Z_t\}$ be a Markov process along some continuous reaction coordinate z, with $Z_0 = z_0$. Let an event be defined by $Z_t = z_{\max}$ for some $z_{\max} > z_0$. Define also the committor probability p that a given realization of $\{Z_t\}$ exceeds $z_{\max}$ before returning to z_0 for $t > 0$, given that $Z_{t>0} \geq z_0$. The event is rare if p is small, namely, $p < 10^{-9}$.

The AMS algorithm begins with the initialization of N replica trajectory segments $\{Z_t^n\}$, $n = 1, \ldots, N$. Simulate the replicas until all of them have returned to z_0 (Fig. 4.1a). Any of these replicas may also exceed $z_{\max}$, at which point the replica is stopped, but the probability is presumably negligible for such an event to occur within N replicas. Obtain the farthest point along z attained by each replica,

$$S_n^1 = \sup(Z_t^n)\,, \tag{4.1}$$

and identify the minimum of these points,

$$q_1 = \min_n(S_n^1)\,. \tag{4.2}$$

Note that at the k^{th} iteration, the proportion of surviving replicas, $1 - 1/N$, provides an estimate $\hat{p}_k$ of the conditional probability that a process starting at z_0 attains a supremum $S > q_k$, given that its supremum is greater than q_{k-1}, i.e. $P(S > q_k | S > q_{k-1})$. In the first iteration, $\hat{p}_1 = 1 - 1/N$ estimates simply the probability $P(S > q_1)$. The probability that a process starting at z_0 exceeds $z_{\max}$ before returning to z_0 is by definition the committor

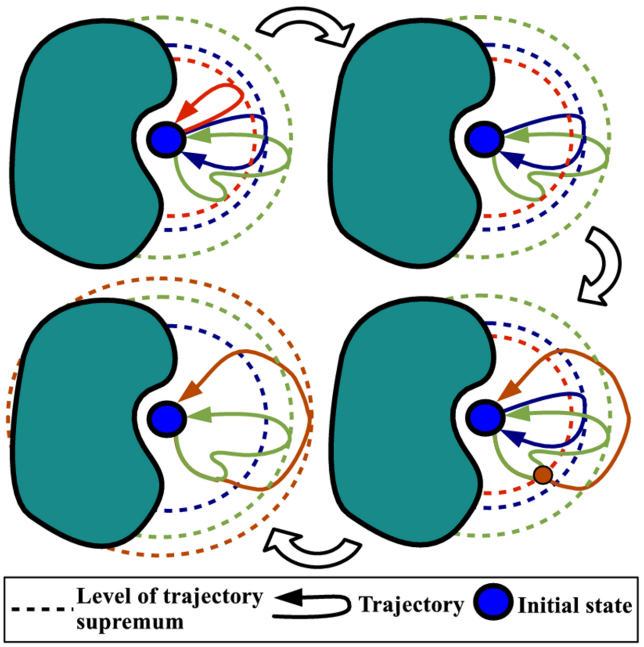

Figure 4.1: Schematic of basic AMS algorithm for a dissociation process of a ligand in an initially bound state from a binding site (blue-green). In this case, $N = 3$ replicas are used and the reaction coordinate is defined as the radius about the initial state. In **(a)**, an initial trajectory segment is generated for each replica. The suprema of the segments are compared, and the replica with the segment of lowest supremum (red) is killed, as shown in **(b)**. Subsequently in **(c)**, a surviving replica is randomly picked (green in this case); its trajectory segment up to the supremum of the killed replica is cloned into the killed replica and simulation is restarted until the trajectory returns to the initial state. Once again, the replica which has the least progress along the reaction coordinate (blue) is identified and killed, as shown in **(d)**. The process is repeated until all replicas have surpassed $z_{\max}$ (not shown).

probability p, given by

$$p = P(S > z_{\max}) = P(S > q_1) \prod_{k=2}^{M-1} P(S > q_k | S > q_{k-1}) \,, \tag{4.3}$$

where M is the number of iterations taken by the AMS algorithm to reach completion.

It can be shown that the product $\hat{p}$ of estimators $\hat{p}_k$ of $P(S > q_k | S > q_{k-1})$ is itself an estimator of the committor probability p [23], achieving equality in the $N \to \infty$ limit, where

$$\hat{p} = \prod_k \hat{p}_k = \left(1 - \frac{1}{N}\right)^M \,. \tag{4.4}$$

The algorithm is here presented for continuous time diffusion process. Slight modifications have to be made for discrete time processes where it is possible for more than one replica to reach the same supremum level [132].

In an idealized setting (namely when the chosen reaction coordinate is the committor function associated with the two sets $A = \{z\,;\, z < z_0\}$ and $B = \{z\,;\, z > z_{max}\}$), it can be shown that the asymptotic variance as $N \to \infty$ is [133, 134]:

$$\mathrm{Var}(\hat{p}) = \frac{-p^2 \log p}{N} \tag{4.5}$$

which can be estimated in practice using $\frac{-\hat{p}^2 log\hat{p}}{N}$, the square root of which provides an estimate of the uncertainty in $\hat{p}$. This estimate should be used with care since it is an asymptotic result and since, in practice, the reaction coordinate is not the committor function. However, it can be used to get a lower bound on the variance. We will discuss in the next paragraph another way to get a safer estimate of the variance.

In the interest of improving efficiency via parallelism, a few variations can be made to the original algorithm described above. Regardless of the choice of N, the mean of $\hat{p}$ is p [132], so that instead of a single AMS simulation with large N, several smaller simulations can be run in parallel to obtain $\hat{p}$ to a similar degree of accuracy through simple averages. Additionally, obtaining multiple estimates of $\hat{p}$ allows for a safe estimation of the variance, by treating $\hat{p}$ itself as a random variable, without having to rely on the need for a large number of replicas through Eq. 4.5. Parallelism may also be incorporated into the algorithm itself, by re-initializing the $(k/N)^{\mathrm{th}}$ quantile at each iteration, killing and restarting $k > 1$ replicas [23, 135]. Nevertheless, it should be noted although the number of iterations can be reduced by using a larger quantile, the variance on the estimator of the probability p will also be larger, and it has been suggested that killing only one replica is the best compromise [136].

4.2.2 Practical Application

The AMS simulations described in the present study run on two different time steps, namely the conventional MD time step and the interval between reaction coordinate measurements, which we term the AMS time step. The MD platform used in the present study, NAMD [137] , does not perform on-the-fly evaluation and comparisons of reaction coordinate values or replica kill-and-restart operations without incurring a large computational overhead. Instead, reaction coordinate evaluation, as well as inter-replica communication and decisions,

are performed at regular time intervals larger than the MD time step in the interest of computational efficiency. This modification does not affect the reliability of the AMS algorithm, which is therefore applied to the subsampled process considered at multiples of the AMS time step. Indeed, as shown Brehier et al [132], the AMS algorithm is unbiased for a discrete-in-time process. A detailed description of the AMS implementation in NAMD is furnished in Section 4.7.

It is also important to note that for physical systems, the initial condition is typically a collection of initial states occupying a continuum on the reaction coordinate, rather than a single value. In a scheme adapted from Cérou *et al* [24], instead of starting the simulation at $z = z_0$, the replicas are initialized and allowed to reach quasi-equilibrium within the defined subspace A of initial states (hereafter called the initial metastable state), assumed to be characterized by the condition $z < z_0$. A value, $z_{\min} > z_0$, is chosen as discussed below, and the replicas are then evolved in time until every replica has reached $z_{\min}$, so as to obtain a representative distribution of trajectories up to the $z = z_{\min}$ hyperplane in configuration space. Thence, the AMS algorithm described above can proceed as shown in Fig. 4.2b. The final state B is similarly defined to be the subspace characterized by $z > z_{\max}$, but this definition does not affect the present implementation of the algorithm.

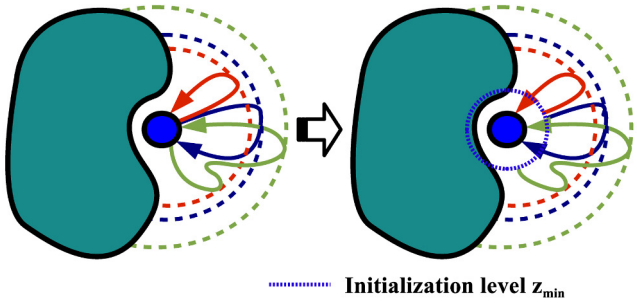

Figure 4.2: Modification of AMS algorithm for efficient simulation. To avoid the difficulty of estimating an extremely small probability, the first step of the original algorithm, shown in **(a)**, is altered to incorporate a starting point $z_{\min}$, shown in **(b)**, a small distance away from z_0. Replica trajectory segments now begin at $z_{\min}$, but still terminate at z_0.

The reason for using initial conditions with a reaction coordinate value greater than z_0 is to avoid the situation where the replicas have a very small probability of reaching B before A, leading to a small estimated committor probability that is typically difficult to estimate accurately. For this reason, $z_{\min}$ should not be chosen too close to z_0. Suitable choices of z_0 and $z_{\min}$ can be determined heuristically from a quasi-equilibrium distribution of the system, such that z_0 and $z_{\min}$ are far enough from each other that the average trajectory from $z_{\min}$

to z_0 is resolvable given the AMS time step. Note that $z_{\min}$ is also bounded above by the requirement that direct simulation can adequately sample trajectory times to and from z_0 to $z_{\min}$ within reasonable computational time.

4.2.3 Calculation of mean first passage time and determination of AMS parameters

The committor probability obtained from AMS provides a means of estimating the mean first passage time from the initial state A to the final state B. For this purpose, model as a geometric distribution the number of non-reactive A-to-A trajectory segments that the system undergoes before a reactive A-to-B trajectory. Thus, the mean of this number can be estimated by the inverse of the measured committor probability, $1/\hat{p}$. To make precise the notion of trajectory loops, define t_1 to be the time taken for a trajectory starting at z_0 to reach $z_{\min}$ and t_2 to be the time taken for a non-reactive trajectory starting at $z_{\min}$ to reach z_0. The expectation of the time taken for one loop is then $\bar{T} = \mathbb{E}(t_1+t_2)$. Additionally, define t_3 to be the time taken by a reactive trajectory, that is one starting at $z_{\min}$ and reaching $z_{\max}$ without first returning to $z_{\min}$. The expected time for a reactive path is then $\mathbb{E}(t_1+t_3)$.

The total time spent by the system in non-reactive trajectory loops is obtained by multiplying the number of trajectory loops by the average time per segment, $\bar{T}$. The time taken for the single reactive trajectory segment at the end then added to the time spent in the the non-reactive loops to produce the AMS estimate of the mean first passage time

$$\hat{\tau} = \frac{\bar{T}}{\hat{p}} + \mathbb{E}(t_1 + t_3)\,. \tag{4.6}$$

An equilibrium simulation, separate from the AMS simulation, is performed to estimate $\bar{T}$. The distribution of loop times is obtained from a projection of the trajectory on the reaction coordinate z, thus providing an estimate of the average loop time and the associated uncertainty. Conveniently, the trajectory also provides the quasi-equilibrium distribution within the initial metastable state, from which the initial AMS replica states can be drawn and suitable values for the parameters z_0 and $z_{\min}$ can be chosen. $z_{\max}$ is chosen by other means, depending on the process being studied. In the present study, a steered MD pulling simulation was employed to obtain a suitable value.

The average time of a reactive trajectory segment, $\mathbb{E}(t_1 + t_3)$ can be obtained from a reconstruction of reactive paths by piecing together the successive trajectory segments tra-

versed by each replica. The resulting collection of paths represent an unbiased distribution of reactive paths, from which the average reactive path time can be obtained. However, it is expected that the mean reactive path time is small compared to the time spent in unreactive loops. This assumption is retrospectively confirmed by the final AMS estimate of the mean first passage time being in at least the millisecond time scale, as compared to the sum of AMS trajectory times, which is in microseconds.

4.3 Simple Validation Test Case

The test system consists of a 50 Å × 50 Å × 50 Å box of explicit water with 0.15 M potassium chloride in solution such that the net charge is zero. One particular K^+ ion is chosen and positioned initially at the origin. Henceforth, the discussion will refer only to this ion. The objective is to evaluate through AMS the characteristic time taken for the ion to migrate from a point of distance z_A from the origin to a point z_B away from the origin, under the influence of a harmonic well potential centered on the origin. For this purpose, the reaction coordinate z is defined to be the distance from the origin r.

CHARMM parameters for ion interactions were taken from Roux and coworkers [138] while the water molecules were characterized by the TIP3P water model [139]. Simulations were run with 1-fs time steps. Long range electrostatic forces were calculated using the particle mesh-Ewald (PME) method with a mesh density of about 1.5 $Å^{-3}$. Van der Waals forces were calculated using a 12 Å cutoff and a switching function starting at 10 Å. Force evaluations were performed at every time step. Periodic boundary conditions were imposed on the faces of the water box and Langevin dynamics was simulated with a temperature of 300 K and damping coefficient of 1 ps^{-1}. Pressure was maintained at 1 atm using a Nosé-Hoover Langevin piston with a damping timescale of 50 fs and a period of 200 fs.

The simulations were carried out in a series of steps. First, the system was energy-minimized over 1000 time steps before being equilibrated for 5 ns with the ion fixed at the origin. Next, $N = 100$ replicas of the system were initialized and run independently, with the ion free to diffuse but under the influence of a spherical harmonic potential $U(r) = \frac{1}{2}kr^2$. Each replica is run until the ion reaches $z_{\min}$ and returns to z_A. This procedure was performed for three different values of k – 0.01 kcal $mol^{-1}Å^{-2}$, 0.02 kcal $mol^{-1}Å^{-2}$, and 0.08 kcal $mol^{-1}Å^{-2}$. After the preparation steps above, the resulting states of the replicas were then fed into three sets of simulations.

The first set of simulations follows the AMS algorithm described above. The committor

Table 4.1: Parameters and results of AMS simulations.

k (kcal mol^{-1}Å^{-2})	z_A(Å)	$z_{\min}$(Å)	z_B(Å)	M	$\hat{p}$
0.01	8	10	22	253	0.079 ± 0.013
0.02	8	12	18	202	0.13 ± 0.02
0.08	5	9	15	474	$(4.5 \pm 1.2) \times 10^{-4}$

probability p estimated by the AMS simulation corresponds to that of the ion, initially at $z_{\min}$, diffusing to z_B without first visiting the sphere $A = \{r : r < z_A\}$. The simulation parameters, number of AMS iterative steps M, and the corresponding $\hat{p}$ values with error estimates given by the square root of the variance are tabulated in Table 4.1.

The second set of simulations consisted of direct 10-ns equilibrium runs on each of the 100 replicas. The long sampling time allowed us to measure the times t_1, t_2, and t_3 by averaging over the times taken for the ion to travel between z_A and $z_{\min}$ and from $z_{\min}$ to z_B. These time values are required to calculate the AMS prediction of τ as per Eq. 4.6. The trajectories obtained also provided direct measurements of τ and p, denoted as $\hat{\tau}_{\text{sim}}$ and $\hat{p}_{\text{sim}}$, respectively. $\hat{\tau}_{\text{sim}}$ is the average of the measured times taken for the ion starting at z_A to reach z_B in the direct simulations. $\hat{p}_{\text{sim}}$ is measured using the formula $\hat{p}_{\text{sim}} = \frac{n(\text{success})}{n(\text{success})+n(\text{fail})}$ where n(success) and n(fail) are, respectively, the number of trajectories starting from $z_{\min}$ that reach z_B before z_A, and the number of trajectories starting from $z_{\min}$ that reach z_A before z_B. With the exception of (*), the aforementioned quantities, listed in Table 4.2, were obtained through direct simulation. (*) was measured from reconstructions of the reactive trajectories generated by the AMS algorithm. In the cases where these direct measurements were possible, the measured p and τ values were compared with those obtained from the AMS runs; however, insufficient "success" events for the $k = 0.08\,\text{kcal mol}^{-1}\text{Å}^{-2}$ case were sampled. Fortunately, an analytic solution is available (see Section 4.4) for comparison in each case, and in particular the latter one.

The analytic model requires the diffusion coefficient of the ion in water, D, to be specified. In the third set of simulations, the local diffusion coefficient was measured at various points in the system for the case $k = 0.08\,\text{kcal mol}^{-1}\text{Å}^{-2}$, through an existing method, which is described together with the results in Section 4.5. In the analytic calculation in Section 4.4, the resulting value of the local diffusion coefficient calculation, $D = 254 \pm 12\,\text{Å}^2/\text{ns}$, was assumed to be constant in space and valid for the other values of k. The uncertainty in τ_{analytic} is due to the uncertainty in D being carried forward.

Theoretical values, p_{analytic} and τ_{analytic}, for p and τ, respectively, were calculated from the

Table 4.2: Direct measurement of variables required for AMS, with the exception of t_3 for $k = 0.08\,\text{kcal mol}^{-1}\text{Å}^{-2}$. The latter was calculated by reconstructing the reactive path trajectories obtained from the AMS algorithm itself. Unexpectedly, it was found for the smallest k value that $T_2 < T_1$ by a small margin. This anomaly is probably due to statistical fluctuations, since the potential in the region $z < z_{\min}$ is almost flat in the small k limit.

k (kcal mol^{-1}Å^{-2})	p_{sim}	t_1 (fs)	t_2 (fs)	t_3 (fs)	τ_{sim} (fs)
0.01	0.092 ± 0.009	35 ± 5	51 ± 8	140 ± 10	940 ± 90
0.02	0.13 ± 0.01	130 ± 10	46 ± 4	60 ± 5	1000 ± 100
0.08	-	340 ± 30	21 ± 1	540 ± 50*	-

Table 4.3: Comparison of p values obtained from AMS, direct simulation, and analytic calculations.

k (kcal mol^{-1}Å^{-2})	$\hat{p}$	$\hat{p}_{\text{sim}}$	p_{analytic}
0.01	0.079 ± 0.013	0.092 ± 0.009	0.084
0.02	0.13 ± 0.02	0.13 ± 0.01	0.13
0.08	$(4.5 \pm 1.2) \times 10^{-4}$	-	3.7×10^{-4}

analytic derivation given in Section 4.4. The estimates from the AMS calculation $\hat{\tau}$ of τ were also obtained, using Eq. 4.6 with values for t_1, t_2 and t_3 from Table 4.2. Tables 4.3 and 4.4 list the results from the AMS calculation, direct simulations, and theoretical calculations for comparison.

Tables 4.3 and 4.4 show that the AMS results compare favorably with those of direct simulation and analytic calculation. Values for p agreed within the error bounds. Minor discrepancies were found in the values of τ, suggesting that the error bounds have been underestimated.

Table 4.4: Comparison of τ values obtained from AMS, direct simulation, and analytic calculations.

k (kcal mol^{-1}Å^{-2})	$\hat{\tau}$ (ns)	$\hat{\tau}_{\text{sim}}$ (ns)	τ_{analytic} (ns)
0.01	1.3 ± 0.3	0.94 ± 0.09	0.99 ± 0.05
0.02	1.5 ± 0.2	1.0 ± 0.1	1.17 ± 0.06
0.08	800 ± 200	-	1040 ± 50

4.4 Analytic derivation of mean first passage time in AMS ion-in-a-well test case

In this section, analytic expressions for p and τ in the ion-in-a-well AMS test case are derived. To recapitulate, the ion is placed in a water box and under the influence of a harmonic well potential of force constant k centered on the origin. Let the time evolution of the ion's position relative to the origin be represented by the stochastic process X_t, with realizations $\mathbf{x} \in \mathbb{R}^3$, obeying

$$\gamma \, dX_t = -kX_t \, dt + \sqrt{2\gamma\beta^{-1}} \, dW_t \, . \tag{4.7}$$

Thus, the process is governed by zero-mass Langevin dynamics with the friction coefficient γ and harmonic potential energy $V(\mathbf{x}) = -k||\mathbf{x}||^2/2$. dW_t denotes a stationary Gaussian process. Recall also that the initial and final states are characterized by the reaction coordinate levels z_A, at the phase space surface ∂A, and z_B, at the phase space surface ∂B, respectively.

Furthermore, let L be the generator of the process X_t, defined for suitable functions f by

$$Lf(\mathbf{x}) = -k\gamma^{-1}x \cdot \nabla f(\mathbf{x}) + (\gamma\beta)^{-1}\Delta f(\mathbf{x}). \tag{4.8}$$

An analytic formula for τ is given as follows. Let $u_0(r)$ be the average time for the process X_t to reach the level z_B, starting at the level r. Then,

$$u_0(r) = D^{-1} \int_r^{z_B} s^{-2} e^{\beta k s^2/2} \left(\int_0^s t^2 e^{-\beta k t^2/2} \, dt \right) ds \, , \tag{4.9}$$

where D is the diffusion coefficient, given by

$$D = (\gamma\beta)^{-1}. \tag{4.10}$$

In particular,

$$u_0(z_A) = \tau, \tag{4.11}$$

where τ is the average time for the process to go from ∂A to B.

The proof of the above formula for τ proceeds as follows. Pick $z \in (0, z_B)$, and let the process X_t be reflected at the level z and absorbed at the level z_B. Define $u_z(\mathbf{x})$ as the average time for X_t to be absorbed, given that $X_0 = \mathbf{x}$ and $z < ||\mathbf{x}|| \leq z_B$. It is known [140]

that $u_z(\mathbf{x})$ is the solution to:

$$\begin{cases} Lu_z(\mathbf{x}) = -1, & \text{if } z < ||\mathbf{x}|| < z_B \\ \nabla u_z(\mathbf{x}) \cdot x = 0, & \text{if } ||\mathbf{x}|| = z \\ u_z(\mathbf{x}) = 0, & \text{if } ||\mathbf{x}|| = z_B \end{cases} \tag{4.12}$$

Putting this equation in spherical coordinates, applying spherical symmetry and using $D = (\gamma\beta)^{-1}$, we get

$$\begin{cases} -D\beta k r u_z'(r) + Dr^{-2}\frac{d}{dr}\left(r^2 u_z'(r)\right) = -1, & \text{if } z < r < z_B \\ u_z'(z) = 0, \quad u_z(z_B) = 1 \end{cases} \tag{4.13}$$

where now u_z has been re-defined as a function of r. Re-writing the above expression gives

$$\left(-\beta k r + 2r^{-1}\right) u_z'(r) + u_z''(r) = -D^{-1}. \tag{4.14}$$

Using the integrating factor $r^2 \exp(-\beta k r^2)$ and the reflective boundary condition we get

$$u_z'(r) = r^{-2} e^{\beta k r^2/2} \int_r^z D^{-1} s^2 e^{-\beta k s^2/2}\, ds. \tag{4.15}$$

Integrating again, using the absorptive boundary condition and finally letting $z \to 0$, we obtain

$$u_0(r) = D^{-1} \int_r^{z_B} s^{-2} e^{\beta k s^2/2} \left(\int_0^s t^2 e^{-\beta k t^2/2}\, dt \right) ds.$$

Next, an analytic expression for p is given. Let $v(r)$ be the probability that the process X_t reaches the level z_B before z_A, starting at the level r. Then

$$v(r) = \left(\int_{z_A}^{z_B} s^{-2} e^{\beta k s^2/2}\, ds \right)^{-1} \int_{z_A}^{r} s^{-2} e^{\beta k s^2/2}\, ds. \tag{4.16}$$

In particular,

$$v(z_{\min}) = p, \tag{4.17}$$

where p is the probability for the process to reach B before A, starting at the level $z_{\min}$.

The proof of the above formula for p proceeds as follows. Let $v(\mathbf{x})$ be the probability for the process X_t to hit the level z_B before z_A, given that $X_0 = \mathbf{x}$ and $z_A \leq ||\mathbf{x}|| \leq z_B$. It is

well known that v is the solution to

$$\begin{cases} Lv(\mathbf{x}) = 0, & \text{if } z_A < ||\mathbf{x}|| < z_B \\ v(\mathbf{x}) = 0, & \text{if } ||\mathbf{x}|| = z_A \\ v(\mathbf{x}) = 1, & \text{if } ||\mathbf{x}|| = z_B \end{cases} \tag{4.18}$$

Using spherical coordinates as above we can re-express the above equations as

$$\begin{cases} -krv'(r) + \beta^{-1} r^{-2} \frac{d}{dr} \left(r^2 v'(r)\right) = 0, & \text{if } z_A < r < z_B \\ v(z_A) = 0, & \text{if } r = z_A \\ v(z_B) = 1, & \text{if } r = z_B \end{cases} \tag{4.19}$$

Thus, through rearrangement of the equation for the case $z_A < r < z_B$,

$$(\beta k r - 2r^{-1}) v'(r) = v''(r)\,, \tag{4.20}$$

which can be integrated to give

$$\log v'(r) = \beta k r^2/2 - 2\log r + C_1\,, \tag{4.21}$$

with C_1 a constant. Finally, we integrate again to obtain

$$v(r) = C_2 \int_0^r s^{-2} e^{\beta k s^2/2}\, ds + C_3\,, \tag{4.22}$$

with C_2, C_3 as constants. Using the boundary conditions to determine C_2 and C_3, we obtain

$$v(r) = \left(\int_{z_A}^{z_B} s^{-2} e^{\beta k s^2/2}\, ds \right)^{-1} \int_{z_A}^{r} s^{-2} e^{\beta k s^2/2}\, ds. \tag{4.23}$$

4.5 Determination of Diffusion Coefficient D for Analytic Model

The analytic model requires the local diffusion coefficient D as one of two parameters. D was measured using a formula due to Woolf and Roux [141] and simplified by Hummer [142], given as follows:

$$D(\mathbf{X} = \langle \mathbf{X} \rangle) = \frac{1}{3} \frac{(\langle \delta\mathbf{X}(t) \cdot \delta\mathbf{X}(t) \rangle)^2}{\int_0^\infty \langle \delta\mathbf{X}(t) \cdot \delta\mathbf{X}(0) \rangle\, dt}\,, \tag{4.24}$$

where $\mathbf{X}$ is the Cartesian coordinates of the ion, $\langle \dots \rangle$ denotes ensemble average (in practice the quantity measured as an average over time) and D is the local diffusion coefficient at position $\langle \mathbf{X} \rangle$.

In accordance with the procedure described in Hummer [142], a potassium ion was allowed to diffuse in the system under the influence of both the harmonic well potential of constant $k = 0.08\,\text{kcal/mol}$, and an additional restraining harmonic potential with constant k_r centered at points of radius $r_0 = 0$, 10, 20 Å away from the origin. It is later found that D does not depend on the local potential gradient, hence it is assumed that the value of D obtained is also valid for other k values. Starting from a state with the ion near r_0, the system was run at equilibrium for 10 ns with data taken every 10-fs interval. Eq. 4.24 was then used to calculate the value of D at the respective points. The runs were repeated for $k_r = 0.1$, 0.3, 0.6 kcal/mol Å^2. The results are as follows:

Table 4.5: Calculation of diffusion coefficient for various physical parameter values.

k_r (kcal mol^{-1}Å^{-2})	r_0 (Å)	D (Å^2/ns)
0.1	0	256
0.1	10	246
0.1	20	245
0.3	0	272
0.3	10	269
0.3	20	256
0.6	0	247
0.6	10	234
0.6	20	258

Taking the mean and standard deviation gives $D = 254 \pm 12\,\text{Å}^2/\text{ns}$.

4.6 Biological Test Case: Benzamidine-Trypsin

4.6.1 The Benzamidine-Trypsin System

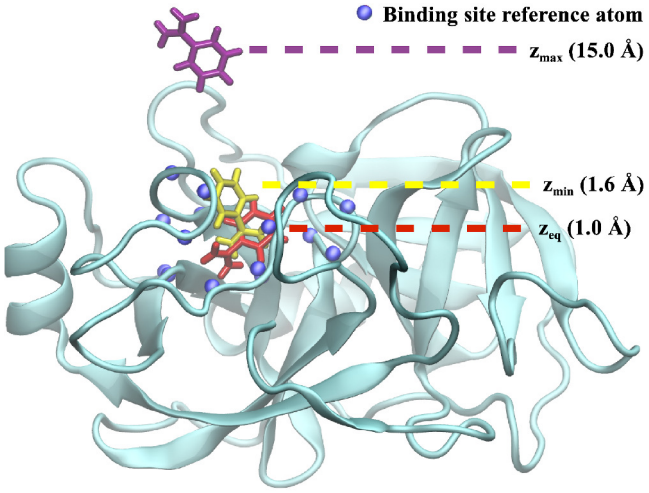

Figure 4.3: Trypsin with various orientations of benzamidine. Trypsin (light blue) and benzamidine in a randomly picked equilibration frame at reaction coordinate z_{eq} (red), AMS initial (yellow) bound states, and AMS final unbound state (purple). The reaction coordinate z is defined as the center-of-mass distance between non-hydrogen benzamidine atoms and the Cα atoms of 16 residues near the binding site (blue spheres). AMS initial and final state structures were extracted from trajectory frames during AMS simulation.

Trypsin is a protease found in many vertebrate species. The complex of trypsin and competitive inhibitor benzamidine is a well studied exemplar of molecular binding and unbinding kinetics [143, 144, 130, 127, 128, 129]. In particular, the rate of dissociation of the bound complex has been measured both experimentally [145] and computationally [127, 128, 129]. The benzamidine-trypsin complex was set up and pre-equilibrated for MD simulation as described in Section S1 of the Supporting Information.

For the AMS simulation, the reaction coordinate z was defined as the center-of-mass distance between Cαs of residues proximal to the binding site (D171, S172, C173, Q174, G175, D176, S177, V191, S192, W193, G194, G196, C197, A198, G204, V205) and benzamidine. The initial (bound) state is characterized by $z < z_0 = 1.6$ Å. z_{min} and z_{max} were chosen to be 1.7 Å and 15 Å respectively, as described in the following section. Fig. 4.3 provides visual examples of benzamidine conformations corresponding to the defined AMS levels, with the structure from a randomly picked equilibration frame displayed for reference.

4.6.2 Determination of AMS Parameters

A 148-ns equilibrium MD simulation of the benzamidine-trypsin complex was performed, with trajectory frames recorded at 0.1-ps time intervals. The projection of the trajectory on z and the normalized distributions of z values are shown in the inset and main figure of Fig. 4.4 respectively.

Parameters for the AMS simulation were chosen as follows. z_0 was set at 1.6 Å, so that the initial state includes the apex of the distribution of the initial bound state, but not too far out so that loop times within the state are kept small and easily sampled. $z_{\min}$ was set at 1.7 Å, close to the value of z_0, but far enough such that the committor probability to be measured was not so small that accuracy is compromised. Initial configurations for the AMS replicas were obtained by randomly drawing frames from the equilibrium simulation that satisfied $z < z_0$.

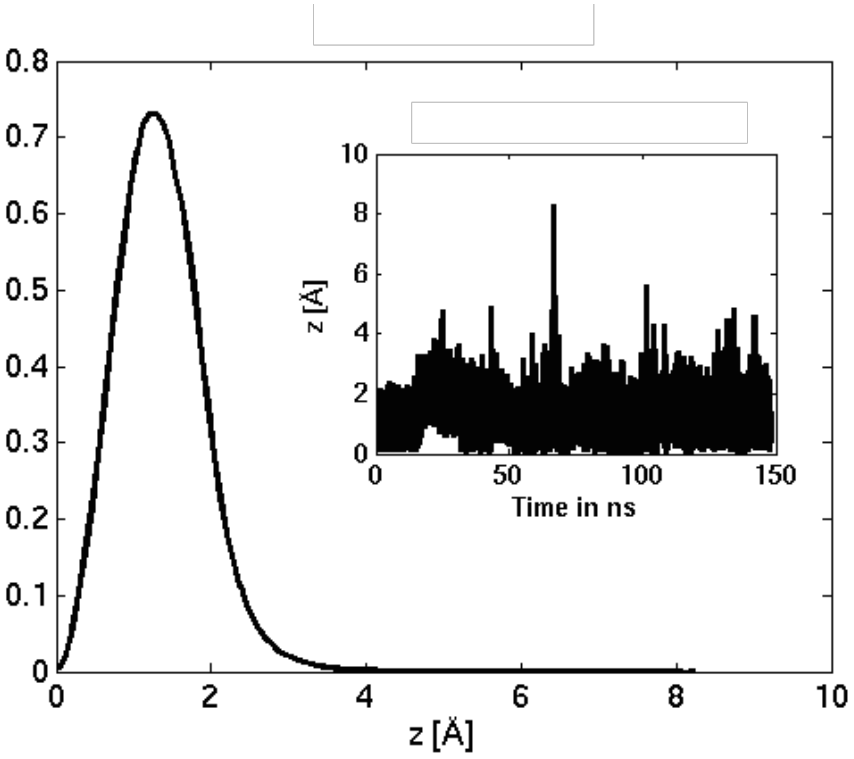

Figure 4.4: Distribution of z values over the course of a 148-ns equilibration. Initial conditions characterizing the bound state, namely the z_0 and $z_{\min}$ levels shown in dashed lines, were chosen such that most of the distribution is included. Shown in the inset is the time series of the z values, from which the distribution was built.

To obtain a suitable value for $z_{\max}$, a constant velocity steered MD simulation [146] was

performed. The average pulling force profile, shown in Fig. 4.5 provides a qualitative survey of the potential of mean force along the reaction coordinate. The force drops to zero at around $z = 13$ Å, thus providing a range at which benzamidine can be considered to have completely dissociated. The endpoint of the simulation, $z_{\max} = 15$ Å, is thus justified.

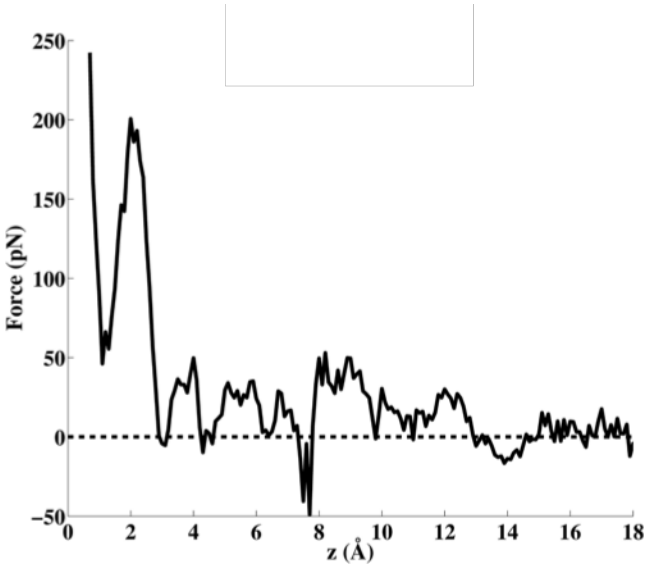

Figure 4.5: SMD force profile. Benzamidine is gradually pulled away from trypsin (details in Section S1 of the Supporting Information). The force profile height reflects the amount of resistance against the pulling force, which drops to near zero when benzamidine is far enough to escape the influence of trypsin. While being a crude measurement of the potential of mean force, this calculation adequately serves as a quick and simple means of locating the unbound state along the reaction coordinate. Note the correspondence of the force peak around $z = 2$ Å to the potential of mean force barrier of the initial metastable state.

4.6.3 Determination of average loop times

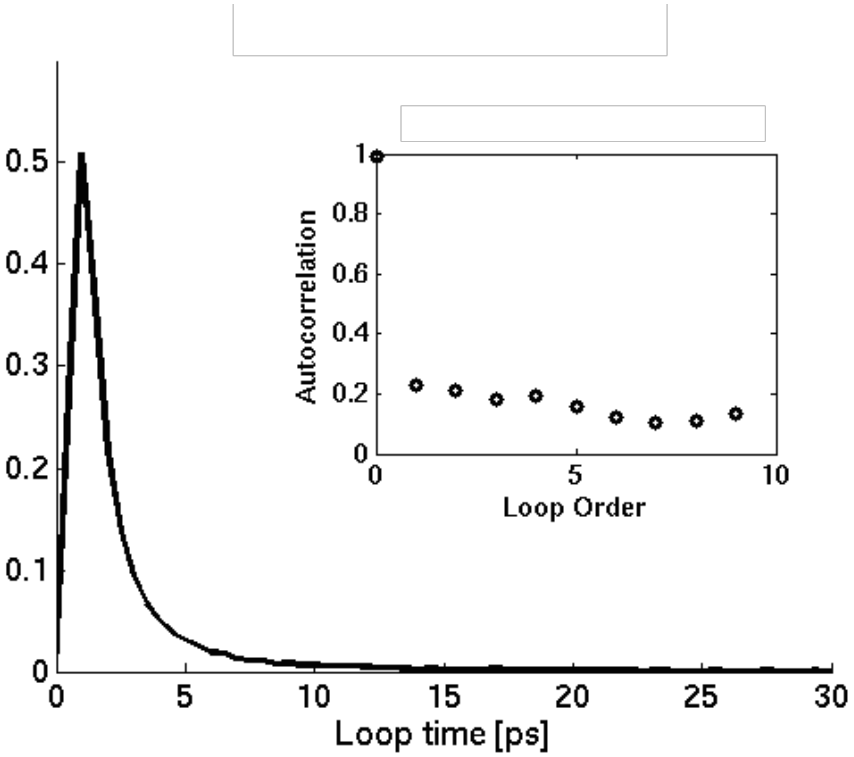

Figure 4.6: Distribution of loop times over a 148-ns equilibration. The average loop time is extracted from the distribution, by fitting to a generalized extreme value distribution, as a necessary step in calculating the dissociation rate. Inset: Autocorrelation function of loop times. Note that the autocorrelation does not fall to zero, indicating that the loop process is not memoryless, as assumed in the AMS derivation. The implications of this non-zero autocorrelation is not clear at this time.

Apart from guiding the choice of starting AMS parameters, the equilibrium simulation was also utilized to estimate the average time taken for the trajectory to loop from $z_{\min}$ to z_0 and back to $z_{\min}$, or equivalently, $t_1 + t_2$. Loop times recorded during the simulation produced the distribution shown in Fig. 4.6, across 36,823 loops.

Fig. 4.6 shows that the distribution has a long tail. To estimate the mean and associated uncertainty, the distribution was fitted to a generalized extreme value (GEV) distribu-

tion [147, 148], yielding maximum likelihood estimates of the parameters

$$\mu = 1.20 \pm 0.01 \text{ ps}, \tag{4.25}$$

$$\sigma = 0.88 \pm 0.01 \text{ ps}, \tag{4.26}$$

$$\xi = 0.71 \pm 0.01, \tag{4.27}$$

where μ, σ, and ξ are the location, scale, and shape parameters respectively, and the uncertainties are given by the 95% confidence interval for each parameter. The corresponding estimate of the mean is then

$$\bar{T} = 3.78 \pm 0.15 \text{ ps}. \tag{4.28}$$

4.6.4 AMS Results

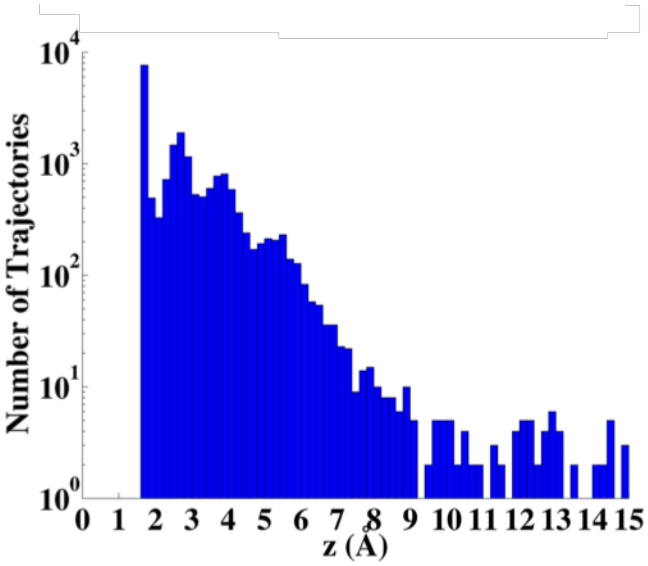

Figure 4.7: Distribution of trajectory starting points. AMS trajectories are histogrammed by branching point z coordinates. Displayed with a logarithmic scale in the y axis, the histogram clearly shows a large concentration of loops about the initial metastable state.

Shown in Fig. 4.7 is a histogram of z-coordinates of branching points during the AMS simulation. A total of 20,376 branching points were required for the 1000 replicas to run to completion. The total simulation time over all replicas was about 2.1 μs. Eqs. 4.4 and 4.5 yield the committor probability estimate $\hat{p} = (5.2 \pm 0.8) \times 10^{-10}$, where the uncertainty is just the root of the variance estimate, given by Eq. 4.5.

Peaks observed at approximately $z = 2.8$, 3.7, and 5.4 Å likely correspond to intermediate metastable states. These states can be problematic, since loops entering these states may linger in them, giving rise to long loop times and consequently the long tail in the loop time distribution. As a result, MD may not be able to sample the loop times adequately. Thus it is recommended that future studies of systems with multiple metastable states apply AMS piecewise between metastable states to extract transition rates for a Markov state model. Nevertheless, it is noted below that the statistical impact from these metastable states is unlikely to be significant in the present case.

Referring again to Fig. 4.7, of all the branching points, 12,579 fell within the range 0 Å to 2.7 Å, roughly corresponding to the primary metastable state. Eq. 4.4 can again be applied to obtain the committor probability of the particle exceeding the heuristic boundary $z = 2.7$ Å of the primary state. This committor probability was found to be $\hat{p'} = 3.4 \times 10^{-6}$. In other words, any loop, on condition that it begins at $z_{\min}$, has only a small probability $\hat{p'}$ of leaving the primary metastable state. Denoting the average loop time of loops originating outside the primary metastable state by $\hat{T}_{\text{hi}}$ and that of loops originating within the primary metastable state by $\hat{T}_{\text{lo}}$, the overall average loop time is given by

$$\bar{T} = \hat{p'}\hat{T}_{\text{hi}} + (1 - \hat{p'})\hat{T}_{\text{lo}}\,. \tag{4.29}$$

$\hat{T}_{\text{hi}}$ is roughly estimated by measuring the mean loop time of all loops that ventured above $z = 2.7$ Å during the equilibration simulation, and was found to be around 250 ps. Assigning $\hat{T}_{\text{lo}}$ the value measured previously in Eq. 4.28, we find that the second term on the left hand side of Eq. 4.29 dominates, so that

$$\bar{T} \approx (1 - \hat{p'})\hat{T}_{\text{lo}} = (3.78 \pm 0.15)\text{ ps}\,. \tag{4.30}$$

The dissociation time estimate is then obtained by applying $\hat{p}$ and $\bar{T}$ to Eq. 4.6 while assuming that the contribution from $\mathbb{E}(t_1 + t_3)$ is negligible. The dissociation time estimate thus obtained is $\hat{\tau} = 0.0075 \pm 0.0014$ s. The corresponding estimated dissociation rates, given by the reciprocal of the estimated dissociation time, is $k_{\text{off}} = 140 \pm 30\ \text{s}^{-1}$, within the same order of magnitude as the experimentally measured rate of $600 \pm 300\ \text{s}^{-1}$ [145]. The overall simulation time taken, summed over all 1000 replicas, was 2.1 μs (2.3 μs after including direct MD and steered MD simulations), which is over three orders of magnitude shorter than the estimated dissociation time of one event.

The dissociation rate estimates obtained in other computational studies of benzamidine-

trypsin [127, 129, 128] differ from the experimental measurement and the present study by between one to two orders of magnitude. However, it should be noted that these studies treated the dissociation process more comprehensively, incorporating multiple distinct bound states that were not considered in the present study. The additional bound states could not have been sampled in the initial equilibration within the AMS initial state, which was shorter than the average transition times between the crystallographic state and the other bound states, reported to be on the order of microseconds [127, 129, 128]. In the same spirit, it should also be noted that the difference in results between the present and reference studies may be due in part to the use of the CHARMM36 [149, 150] force field in the present study, in contrast to the AMBER force fields employed in the other studies.

4.7 Software Implementation

The AMS test cases were simulated using a prototype implementation of the AMS algorithm on NAMD. The main issues addressed in this implementation were simulation scalability, data management, and method parallelization. A general outline of the AMS implementation and organization is shown schematically in Fig. 4.8. Under this scheme, the AMS control code works in conjunction with NAMD to set up, run, and analyze each simulation step. A pool of NAMD instances utilize a shared file system to allow cross-process communication, enabling dynamic initialization and termination of simulations as guided by the AMS control logic.

As the centerpiece of the described implementation, the NAMD software package [137] was chosen for its highly efficient, scalable, and feature-rich MD engine that can run on a variety of platforms and is maintained at most supercomputing centers around the world. Two NAMD features of primary importance to AMS are the native "Colvars" implementation [151] and the embedded Tcl interpreter. The colvars module is leveraged to define and monitor the AMS reaction coordinate using the built-in collective variables, which are easily defined and can be combined to describe complex reaction coordinates. Access to the Tcl scripting interface, a feature unique to NAMD, allows a great deal of flexibility to rapidly develop and debug the AMS control code without requiring detailed knowledge of the internal workings of NAMD. Separating the AMS control from the NAMD "black box" requires only minor modifications to the NAMD source code–adding mechanisms for reading/writing of restart data and sequential trajectory files. Although these features were added specifically to enable the AMS method as described herein, they are of general utility to NAMD users and have

been included in NAMD since the version 2.10 software release.

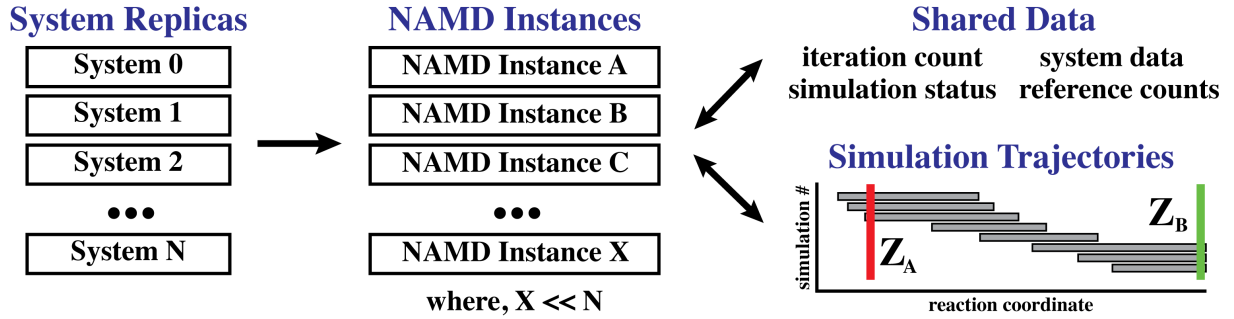

Figure 4.8: Schematic of AMS implementation. The AMS control logic running within each NAMD instance monitors global progress using a shared file system to communicate information between processes. Upon selecting a system replica for restart, the associated positional, velocity, and periodic cell data is loaded into an available NAMD instance and the simulation is launched. During the course of each simulation, NAMD continually saves the simulation trajectory to disk and updates multiple elements of shared data used to drive the AMS control logic.

From the outset, coupling the AMS methodology to MD simulations raised significant concerns regarding data management; MD is data-intensive (generates large positional trajectories and restart files), while AMS is highly duplicative (reactive trajectories are "copied" up to the branch point). Accordingly, the storage requirements of an AMS run are mitigated using two concurrent approaches. The first approach is centered on minimizing the number of simulation restart files that are stored. As described in the Branching Step of the algorithm, simulations are restarted at the *first* crossing of a particular level of the reaction coordinate. In practical terms, this specification restricts potential restarts to frames in which the z-value is a new maximum observed value, as shown in Fig. 4.9. Each set of restart data (position, velocity, periodic cell) is stored using the frame number, preserving the timing information of the simulation. The largest reduction in storage requirements can be realized by suppressing the positional trajectory output (DCD files) without compromising the calculation of p and τ, albeit at the loss of continuity in the atomic detail for each reactive trajectory.

The second approach employs reference counting, a computer science framework for managing objects in memory, to curate the simulation data. As implied in Fig. 4.8, the simulation data (trajectories, restart data) are stored and manipulated as a collection of individual files. Each system replica maintains an ordered list of references to frames within a particular simulation file that, when stitched together, defines the reactive trajectory. The number of

references to each file are maintained as part of the shared data, and reference counts are incremented or decremented as trajectories are duplicated or discarded, respectively. When the reference count for a particular file reaches zero, the data is no longer relevant to any of the surviving reactive trajectories, and therefore, deleted from the file system. Using this framework, duplicate data is minimized and simulation data that is no longer relevant to the remaining reactive trajectories is immediately discarded.

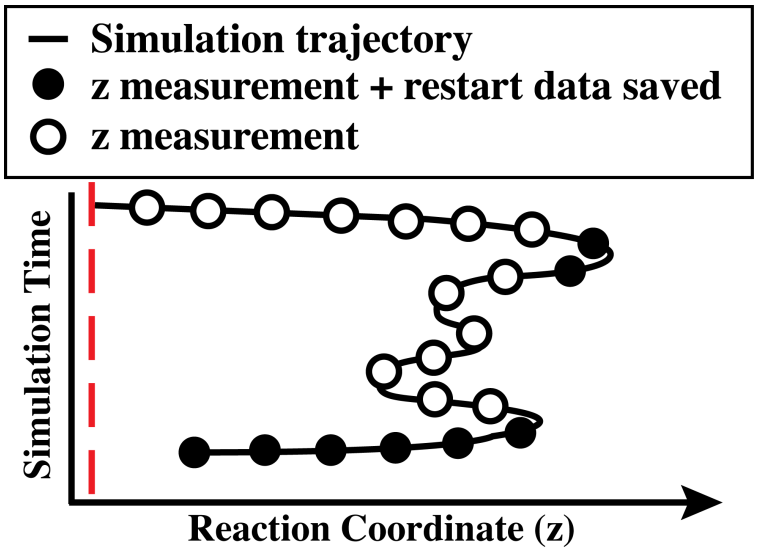

Figure 4.9: Illustration of a typical simulation trajectory. Each circle represents a point at which the reaction coordinate z is measured. In the interest to data economy, restart data is only written to memory when the measured reaction coordinate achieves a new global maximum (filled circles), representing a valid restart point for subsequent simulations.

Finally, a critical component of the present implementation is the "pseudo-parallelization" of the AMS method. In the basic description of the AMS algorithm, while one replica runs, the AMS system waits for the current simulation to reach a termination criterion (return to state A or advance to state B) before starting a new replica (as in Step 6 of the algorithm). In the present implementation, we instead start a new replica as soon as the smallest maximum level among all running replicas surpasses the smallest maximum level among all stopped replicas. Thus, many replicas can be running concurrently in parallel. Fig. 4.10 depicts this strategy in which the least advanced trajectory, shown in green, represents a currently running replica. Once this replica has surpassed the threshold demarcated by the orange line, the least-progressed stopped replica (blue), can be restarted. Although the degree of parallelism and replica start times are unpredictable (an outcome of the stochastic MD process), this design allows for significant parallelization in two key areas: as simulations

rapidly progress along the reaction coordinate in areas of low energy, and when simulations make any degree of progress along the reaction coordinate (*e.g.*, high energy) but require a non-trivial amount of time to return to state A.

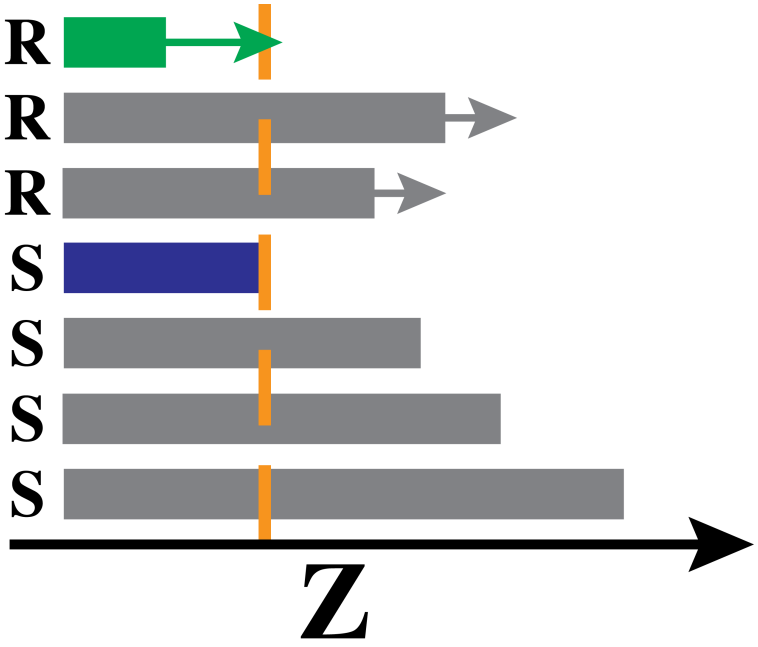

Figure 4.10: Illustration of "pseudo-parallelization" using a chart of maximum z attained in an example AMS simulation with seven replicas. The letters "R" and "S" label running and stopped replicas, respectively. When all running simulations have surpassed the level of the least-progressed stopped replica (shown in blue), an event denoted by the green replica crossing the orange threshold, the blue-colored replica can begin running even before any running replicas terminate.

4.8 Conclusion

The results of the present study demonstrate the potential of AMS in problems involving rare event sampling in MD. In the simple idealized test cases and benzamidine-trypsin dissociation problem studied here, AMS was used successfully to estimate the committor probabilities and the mean first passage times of the processes studied, achieving agreement with direct MD simulation and theoretical calculations in the simple test cases, and order-of-magnitude agreement of k_{off} with experiment in the case of benzamidine-trypsin. For comparison, other

methods to calculate the benzamidine-trypsin dissociation rate through MD [127, 129, 128] have produced results that were between 2 to 4 orders of magnitude off.

Despite the success of AMS demonstrated so far, significant sources of error exist in the methodology, especially in the benzamidine-trypsin test case. Thus it cannot be ruled out that the agreement of k_{off} with the experimental value is due to a serendipitous cancellation of errors. For example, the duration of a given loop is not completely independent of the loop that came before it, or even to loops that occurred several instances before, as shown by the inset of Fig. 4.6. The impact of this memory effect has not been quantified, but is currently under investigation. Other sources of error originate from the complexity of the system being studied. One such source of error was the omission of alternative initial bound states, as discussed in Results and Discussion. A related problem is the existence of intermediate metastable states which, although detected during the AMS run, were not extensively sampled. Although it was argued that the impact of the overlap was small, measures to prevent such overlaps should nevertheless be undertaken in future AMS studies of complex systems. These problems were not present in the simple test cases, and might explain why better agreement with the benchmarks was achieved for them than in the benzamidine-trypsin case. Before AMS can be adopted for mainstream use, the problems described above have to be addressed. A number of remedies are prescribed below for this purpose.

One solution is to choose a reaction coordinate and initial state z value that would encapsulate all metastable states, if they are close enough to one another. For this purpose, visual inspection of the equilibrium simulation is recommended as a good practice for AMS methodology. In cases where the metastable bound states can be characterized, detailed information about the system can be obtained by applying AMS piecewise to determine transition rates between the various states of the system in a Markov state model, in similar fashion to the other studies referenced [127, 129, 128].

It should also be noted that unaccounted-for sources of error, inherent in MD-based methods, were not incorporated into the uncertainty of the dissociation rate estimate. In particular, the use of non-polarizable force fields may significantly affect the interaction strength between ligand and substrate. Tiwary *et al* [128] suggest that the use of a non-polarizable force field results in an overestimate of the dissociation time. Polarizable force fields such as the Drude model [152, 153] may be used to obtain more accurate descriptions of dissociation dynamics where solvation and other polarization-dependent effects can be foreseen to play a major role.

Nonetheless, the favorable results reported in the present study merit further developmental effort to address the present inadequacies and improve the technique; experimental evidence suggests that significant variation in drug efficacy occur with within-order-of-magnitude differences in residence times. For example, a study of a set of inhibitor compounds of the FabI enoyl reductase in mice infected with *Francisella tularensis* showed an approximate 1.2% increase in survival rate for each 1-min increase in residence time in the range of 20 to 140 mins [154]. In a study of A_{2A} receptor agonists, an almost twofold increase in efficacy was found for each increase in order of magnitude of residence time over the 1- to 100-minute range. [155] In order to further test the applicability of AMS in this respect, future efforts are required to address the challenges of initial state definition and loop time sampling encountered in the present study, as well as to validate AMS in other molecular systems. Various technical issues also need to be resolved, such as improving time resolution by enabling more frequent reaction coordinate queries, efficient communication between replicas, and increased parallelism of the algorithm.

CHAPTER 5

KINETIC MODEL OF MOLECULAR DIFFUSION[1]

In this chapter, an algorithm for the simulation of molecular solutes in complex, highly detailed environments is proposed. The biological context of the problem is in biomolecular transport processes, which often involve solute dynamics in the vicinity of macromolecules, like membrane channels, as well as diffusive approach of solutes to the macromolecules. The latter occur on length scales of between 10 to 100 nm, and is influenced by a highly particular environment, constituted of macromolecular geometry and the surrounding electrostatic field. The proposed algorithm describes solute energetics and mobility in such an environment through a kinetic model of diffusion based on a Markov state model framework. Prerequisite input data consist of diffusion coefficient and potential of mean force maps generated from extensive molecular dynamics simulations of proteins and their environment that sample multi-nanosecond durations. The suggested diffusion model can describe transport processes beyond microsecond duration, relevant for biological function and beyond the realm of molecular dynamics simulation. The system being simulated is represented by a discrete set of states corresponding to cells in a Voronoi tessellation of the system, distributed according to a density function that resolves intricate regions of the diffusion space to a sufficient level of detail. Each state is specified by the position, volume, and surface elements of the corresponding Voronoi cell. Simple validation test cases demonstrate that the model and the associated Brownian motion algorithm are viable over a large range of parameter values such as time step, diffusion coefficient, and grid density. Two biological applications are also described. The first application is the translocation of a nascent protein chain from the translocon interior to the exterior lipid environment, and the second is ion diffusion around and through the *Eschericia coli* mechanosensitive channel of small conductance ecMscS.

[1]The research presented in this chapter has been published in I. Teo and K. Schulten, *J. Chem. Phys.*, **139** (2013), 121929, and J. C. Gumbart, I. Teo, B. Roux, and K. Schulten, *J. Am. Chem. Soc.*, **135**(6) (2013), p. 2291–2297. .

5.1 Introduction

Diffusion is a mainstay of biological systems across many time and length scales. On the biological cell level, many phenomena have been framed as diffusion-controlled processes, from transport processes [156, 157, 158], ligand binding [159, 160, 161, 162, 163] and signal transduction within the cell [164, 165, 166, 167], to cell-to-cell signaling [168, 169, 167]. These processes can depend on molecular-level detail in regard to the geometry of the diffusion space, energetics and local variation of diffusivity. Experimental investigations of molecular-scale transport are often unfeasible. Fortunately, observations can be complemented by computer simulations. In fact, diffusion theory [157, 170, 161] is well-established, making diffusion-controlled processes amenable to computer simulation. However, most applications of diffusion theory in the past have glossed over the molecular-scale variation of geometries, energetics, and mobilities of transported solutes.

Biophysics has made great progress in understanding the regulation of transport at the intra-protein level, particularly in the case of membrane channels. However the intra-protein steps are preceded by diffusive approach and control of access to the relevant surface openings of channel proteins, in particular, since the relevant overall diffusion space is often highly intricate in regard to local geometry as well as solute energetics and mobility. Spatial and time scales for the diffusive approach are typically 10-100 nm and 1 ms, respectively. In the present study we suggest and test a flexible computational scheme to describe the initial diffusive approach step of biological transport. This scheme is based on extensive prior sampling through nanosecond molecular dynamics (MD) simulations and a subsequent application of diffusion theory that furnishes extremely realistic microsecond to millisecond descriptions at molecular resolution.

A diverse set of simulation techniques are already routinely employed to model diffusion on the spatial and temporal scales relevant for cellular transport mechanisms. Programs such as Smoldyn [171], MCell [172] and VCell [173] have been used successfully for reaction-diffusion simulations. but on a larger scale than considered in the present study, namely on the scale of whole cells. Another approach commonly used to describe biological diffusion processes is that of Green function reaction dynamics [37, 38, 39], which solves the diffusion equation for one particle or two particles and uses the resulting Green function solution to propagate particle positions in time. Until now, the aforementioned computational tools assume free diffusion or the presence of a simple potential, typically arising from a few inter-particle interactions, and in this case are able to describe large systems well. On the much smaller molecular scale, however, inter-molecular interactions with the environment need to

be accounted for through detailed, complex potentials that require descriptions based on advanced numerical techniques. Of these techniques, MD remains the most detailed, but also the computationally most expensive technique, the expense placing limits on spatial and temporal scales that can actually be covered [174, 175, 176]. To overcome such limitations, MD must often be supplemented by sampling techniques and parallelization schemes [177]. Brownian dynamics (BD) [32], which sacrifices some level of detail by treating solvents as implicit and large molecules as reflective barriers without internal degrees of freedom, has been successfully used to simulate larger systems [33, 34, 35]. Efforts to extend the reach of molecular-level simulations to greater length and time scales include diffusion Monte Carlo algorithms, such as that implemented in BioMOCA [36], and mean field descriptions of diffusion, commonly implemented through a finite element approach [169, 165, 166]. However, particulate detail may be required in certain cases, such as high proximity interactions in the narrow diffusion space within ion channels.

In any computational investigation of a diffusive process, the limitations of existing simulation methods described above necessitates a careful choice of method, guided by the scale of the system, knowledge of the process to be studied, and the availability of computational resources. The latter two factors, together with the need to be familiar with multiple simulation methods, present a hurdle to cross during early-stage investigations. Obtaining enough sampling of diffusive processes on relevant time scales presents a challenge to atomistic methods like MD and BD. In the particular case of MD, one may not observe an expected phenomenon even after extensive sampling. On the other hand, coarse-grained simulations may not be adequately detailed to reproduce diffusive motions influenced by intricate environmental effects. Thus, there is a motivation to interface MD calculations with coarse-grained diffusion algorithms to take advantage of the atomic detail obtained by the former and the long time scales accessible to the latter. One such approach is atomic-resolution BD, a variant of BD that uses MD-derived potentials of mean force (PMF) and diffusivities to describe interactions of diffusing particles with each other and with surfaces [40]. These PMF and diffusivity maps are obtained from relatively short all-atom MD simulations using advanced sampling techniques. Atomic-resolution BD has been shown to reproduce the results of MD simulations and has been used successfully in a number of applications [178, 179, 40].

The proposed algorithm capitalizes on the finite element methodology, which has the ability to describe arbitrary potential fields and local mobilities, as in the case of atomic-resolution BD. The method offers the flexibility of a multi-resolution grid and permits descriptions afforded by the MSM protocol [131] to develop a versatile particle-based method

that is valid and computationally viable over a wider range of length and time scales as compared to other diffusion methods, without compromising the level of detail in describing the potential field. Through the MSM scheme, one can divide the computational effort in an extremely useful manner between a sampling step that gathers the physical characteristics of the diffusion model and a diffusion execution step that describes the actual transport between cell environment and protein channel. This division allows the description of diffusion within an arbitrary potential field, extending earlier methods applicable in case of large systems only to free diffusion.

The need for detail in large systems is motivated by the aim of describing the diffusive approach of solutes in simulations of membrane channels, as discussed further in Section 5.6. Our algorithm models the system at varying levels of detail and optimizes computational efficiency through adjustment to the level of detail required for the system's description. Our method allows also the use of large time steps, extending thereby the reach of simulations to time scales longer than those of other molecular-scale methods.

In our algorithm, diffusion is implemented through a kinetic model of particles transitioning between pre-defined states. The rates of transition between states are specified by a rate matrix. Given the size of a time step, one can obtain the respective probabilities of the particle transitioning from its current state to each of the other states within the span of a time step by solving for the eigenvalues and eigenvectors of the rate matrix.

The set of states is characterized by positions in the system, namely the centers of cells in an irregular grid of varying resolution overlaid on the system. The rate matrix is calculated from the discretized Smoluchowski equation, using pre-obtained input parameters, namely the diffusion coefficient and a potential of mean force (PMF) map of the system for the diffusing species. Solving the eigenproblem of the rate matrix then gives the transition matrix, which propagates a particle's position through time. Key to the method is the use of prior MD simulations to extract the diffusion coefficient and PMF map of the system.

The cells are obtained by adapting the distribution of a set of points to an input density function which attributes a higher density to regions of the system of more intricate geometry, such as near the surface of a macromolecule, and performing a subsequent Voronoi tessellation. Each cell is geometrically defined through position, volume, as well as number and sizes of surface elements shared with neighboring cells. As stated in the previous paragraph, the physical characteristics of the cells, namely the PMF and diffusion coefficient, are gathered through extensive MD simulation.

The implementation of the finite difference scheme for diffusion calculations has been

carried out at this point in Matlab [180]. The subroutines used and specified below were taken from existing Matlab libraries.

For the purpose of validation, the proposed algorithm has been used to reproduce the expected statistical behavior in simple systems for which analytic descriptions are known. It was found that, to the extent of the values of parameters tested (namely time step and diffusion coefficient), the accuracy of the scheme is limited by the resolution of the grid used to discretize the system. The model is subsequently applied to realistic cases, the first involving diffusion of a nascent protein chain in the 2-dimensional plane of a lipid membrane out of a translocon, and the second involving ion diffusion through the *Eschericia coli* mechanosensitive channel of small conductance (ecMscS).

5.2 Model Building and Simulation Algorithm

Actual use of the methodology suggested in the present study begins with MD simulations sampling solute energetics and diffusion coefficient maps. These maps serve as input into the diffusion algorithm. However, the diffusion algorithm will be presented first in this section, and subsequently applied to test cases that do not require MD simulation input. In the biological applications (Sections 5.5 and 5.6), MD simulations will be used to derive the required inputs prior to running the diffusion algorithm. The reader should note that in real use cases, the MD simulations are always carried out first.

Upon the premise that the MD simulations have been run and the relevant data obtained, we describe the derivation of the algorithms involved in the diffusion model. The algorithms employed in setting up the geometrical and physical details of the diffusion space are described first. The following section will formulate the discrete form of the Smoluchowski equation governing transport processes on time scales of picoseconds or longer, i.e., in the strong friction regime. Finally, the numerical solution employed for the BD simulation is described.

5.2.1 Setting Up a Discrete Representation of the System

The efficiency of the diffusion algorithm is dependent on the choice of grid used to discretize the system. In this regard, one should choose a grid density profile suited to the local level of detail in the description of the system, while minimizing the error arising naturally from approximating a continuous space with a discrete one. Our approach employs for this purpose

a topology-conforming self-organizing map [181, 182, 183] that distributes cell centers and determines then the respective Voronoi tessellation.

Let X, a subset of $\mathbb{R}^n$, represent the diffusion space. Typically, X is a subset of the three-dimensional space $\mathbb{R}^3$. $U(\mathbf{x})$ and $p(\mathbf{x},t)$ are, respectively, the underlying potential felt by a particle at $\mathbf{x} \in X$ and the probability distribution of diffusing particles at $\mathbf{x}$ and time t. Our aim is to divide the otherwise continuous system up into a discrete collection of regions. For this purpose, a set of N points $\{\mathbf{w}_i \in X \mid i = 1, 2, \ldots, N\}$ are selected at random. These points represent the centers of cells in the Voronoi tessellation grid $\mathcal{I}$ of X. Formally, $\mathcal{I}$ is defined through

$$\mathcal{I} = \Bigg\{ I_i \subset X \;\Bigg|\; \bigcup_i I_i = X \text{ and } I_i \cap I_j = \emptyset \;\forall\; i \neq j \;,\; \mathbf{x} \in I_i \iff ||\mathbf{x} - \mathbf{w}_i|| < ||\mathbf{x} - \mathbf{w}_j|| \;\forall\; i \neq j \;,\; i = 1, 2, \ldots, N \Bigg\} \quad . \tag{5.1}$$

Before constructing $\mathcal{I}$, we adapt the positions $\{\mathbf{w}_i\}$ to conform topologically to the predefined distribution $\rho(\mathbf{x})$. The resulting grid will have two desirable properties.
First, the density of cells is locally homogeneous, i.e., the property holds

$$r \to r_\rho(\mathbf{x}) \implies \rho_w(S_r(\mathbf{x})) \to \rho(\mathbf{x}) \quad , \tag{5.2}$$

where $\rho_w(S_r(\mathbf{x}))$ is the density of cell centers $\mathbf{w}_i$, and thus of cells, within a sphere $S_r(\mathbf{x})$ of radius r centered on $\mathbf{x}$, and $r_\rho(\mathbf{x})$ is the length scale associated with $\rho(\mathbf{x})$, so that $r_\rho(\mathbf{x}) \sim \rho(\mathbf{x})^{-1/3}$. Thus the local diffusive behavior of particles will be subject to minimal error arising from local grid density deviations from $\rho(\mathbf{x})$.
Second, $\mathcal{I}$ is a centroidal Voronoi tessellation (CVT) [184]. In other words, each $\mathbf{w}_i$ coincides with the centroid of cell I_i. The CVT property minimizes the mean square deviation of position (MSD) within I_i, given by

$$\mathrm{MSD}_x^{(i)} = \left(\int_{I_i} d\mathbf{x} \; ||\mathbf{x} - \mathbf{w}_i||^2 \right) \Big/ \int_{I_i} d\mathbf{x} \quad . \tag{5.3}$$

Under the assumption that $p(\mathbf{x})|_{I_i}$ is, on average, locally monotonic with respect to $||\mathbf{x} - \mathbf{w}_i||$, the chosen discretization scheme also minimizes the global mean square error in p, given by

$$\sum_{i=1}^{N} \mathrm{MSD}_p^{(i)} = \sum_{i=1}^{N} \left(\int_{I_i} d\mathbf{x} \; |p(\mathbf{x}) - p(\mathbf{w}_i)|^2 \right) \Big/ \int_{I_i} d\mathbf{x} \quad . \tag{5.4}$$

A topology-conforming distribution of $\{\mathbf{w}_i\}$ can be constructed using an iterative procedure, due to Martinetz, Berkovich and Schulten [182], comprised of the following four steps:
1. Begin with any random distribution of $\{\mathbf{w}_i\}$ over X.
2. From the reference distribution $\rho(\mathbf{x})$, draw a test point $\mathbf{v} \in X$.
3. Rank $\{\mathbf{w}_i\}$ according to each respective element's distance from $\mathbf{v}$. According to this order, assign a rank integer $k(\mathbf{v}, \mathbf{w}_i) = 0, 1, 2, \ldots, N-1$ to each $\mathbf{w}_i$, increasing from the nearest to the farthest.
4. Update each $\mathbf{w}_i$ as follows:

$$\mathbf{w}_i(s+1) = \mathbf{w}_i(s) + \epsilon(s)(\mathbf{v}(s) - \mathbf{w}_i(s))e^{[-k(\mathbf{v}(s),\mathbf{w}_i(s))/\lambda(s)]} \quad , \tag{5.5}$$

where s labels the adaptation step, ϵ and λ control, respectively, the magnitude of the change and the extent of the area of influence around $\mathbf{v}$. Each adaptation step s comprises steps 2 to 4. The prescription [182] calls for a gradual decrease of ϵ and λ, such that for each s,

$$\epsilon(s) = \epsilon_{\text{initial}} \left(\frac{\epsilon_{\text{final}}}{\epsilon_{\text{initial}}} \right)^{s/S} \quad , \tag{5.6}$$

$$\lambda(s) = \lambda_{\text{initial}} \left(\frac{\lambda_{\text{final}}}{\lambda_{\text{initial}}} \right)^{s/S} \quad , \tag{5.7}$$

where S is the total number of adaptation steps chosen by the user. Thence, set $\mathbf{w}_i \equiv \mathbf{w}_i(S)$. The parameters $\epsilon_{\text{initial}}$, ϵ_{final}, λ_{initial}, λ_{final} and S must be tuned through trial-and-error to achieve convergence to the desired distribution of $\mathbf{w}_i$. Convergence may be characterized using the mean square deviation of the grid or of some subset of cells J in the grid, namely $\sum_{i \in J} \text{MSD}_x^{(i)}$.

Having determined $\mathbf{w}_i$, the Matlab subroutine `DelaunayTri` is employed to extract the Delaunay triangulation. The latter is obtained by connecting lines between pairs of nearest neighbor $\mathbf{w}_i$'s. The Voronoi tessellation is calculated from the Delaunay triangulation by means of the Matlab subroutine `voronoiDiagram`. The tessellation is specified by the vertices of each cell I_i, from which the cell volumes and interfacial cell surface areas needed for the discretization of the diffusion equation can be calculated.

5.2.2 Discretization of the Smoluchowski equation

In the presence of a potential $U(\mathbf{x})$, diffusion is described by the Smoluchowski equation in continuous space:

$$\dot{p}(\mathbf{x},t) = \nabla \cdot D(\mathbf{x}) e^{-\beta U(\mathbf{x})} \nabla e^{\beta U(\mathbf{x})} p(\mathbf{x},t) \quad , \tag{5.8}$$

where $D(\mathbf{x})$ is the local diffusion coefficient and $\beta = (k_{\mathrm{B}}T)^{-1}$. This equation had been used extensively to describe diffusion outside and inside of the ecMscS cytoplasmic domain [158].

In order to choose $D(\mathbf{x})$ and $U(\mathbf{x})$, a MD simulation under equilibrium conditions can be used to calculate the diffusion coefficient and generate a PMF map, which is subsequently interpolated to provide an associated potential value for every grid cell. Such analysis is demonstrated for MscS in Section 5.6.1. In case that simplified models for $D(\mathbf{x})$ and $U(\mathbf{x})$ suffice, the two quantities can be obtained by some other means, e.g., as in a recent study of SecY [185], where the potential map within a protein channel was obtained by calculating the potential along a specific radial direction and generalizing the result by assuming radial symmetry.

With the required data in hand, we discretize Eq. (5.8) in space by integrating over a generic cell I_i resulting in

$$\int_{I_i} d\mathbf{x}\ \dot{p}(\mathbf{x},t) = \int_{I_i} d\mathbf{x}\ \nabla \cdot D(\mathbf{x}) e^{-\beta U(\mathbf{x})} \nabla e^{\beta U(\mathbf{x})} p(\mathbf{x},t) \tag{5.9a}$$

$$= \int_{\partial I_i} d\sigma\ \hat{\mathbf{n}}(\mathbf{x}) \cdot D(\mathbf{x}) e^{-\beta U(\mathbf{x})} \nabla e^{\beta U(\mathbf{x})} p(\mathbf{x},t) \tag{5.9b}$$

$$= \int_{\partial I_i} d\sigma\ D(\mathbf{x}) e^{-\beta U(\mathbf{x})} \frac{\partial}{\partial \hat{\mathbf{n}}} e^{\beta U(\mathbf{x})} p(\mathbf{x},t) \quad . \tag{5.9c}$$

Here Gauss' theorem has been applied in the second line to obtain an integral over the surface ∂I_i of cell I_i with $\hat{\mathbf{n}}(\mathbf{x})$ representing the unit surface normal. The dot product in the second line is then converted to a directional differential in the third line.

Next, we make the approximation that the quantities $p(\mathbf{x},t)$, $U(\mathbf{x})$ and $D(\mathbf{x})$ are uniformly valued in the interior of each cell i with center $\mathbf{w}_i$ resulting in

$$p(\mathbf{x},t) \approx p(\mathbf{w}_i,t)\ \forall\ \mathbf{x} \in I_i \backslash \partial I_i \quad , \tag{5.10}$$

$$U(\mathbf{x}) \approx U(\mathbf{w}_i)\ \forall\ \mathbf{x} \in I_i \backslash \partial I_i \quad , \tag{5.11}$$

$$D(\mathbf{x}) \approx D(\mathbf{w}_i)\ \forall\ \mathbf{x} \in I_i \backslash \partial I_i \quad . \tag{5.12}$$

Furthermore, we set the values of variables at each cell interface to be the average of the

values in the two cells, namely

$$De^{-\beta U(\mathbf{x})}\frac{\partial}{\partial \hat{\mathbf{n}}}e^{\beta U(\mathbf{x})}p(\mathbf{x},t)\bigg|_{\partial I_{ij}} \approx \frac{D(\mathbf{w}_i)+D(\mathbf{w}_j)}{2}\; e^{-\beta[U(\mathbf{w}_i)+U(\mathbf{w}_j)]/2}\;\times \\ \frac{e^{\beta U(\mathbf{w}_j)}p(\mathbf{w}_j,t)-e^{\beta U(\mathbf{w}_i)}p(\mathbf{w}_i,t)}{||\mathbf{w}_j-\mathbf{w}_i||} \quad , \tag{5.13}$$

where ∂I_{ij} is the interface between I_i and I_j. Putting Eqs. (5.12,5.13) into Eq. (5.9c) gives

$$V_i\;\dot{p}(\mathbf{w}_i,t) = \sum_{j\neq i} A_{ij}\frac{D(\mathbf{w}_i)+D(\mathbf{w}_j)}{2}\; e^{-\beta[U(\mathbf{w}_i)+U(\mathbf{w}_j)]/2}\;\times \\ \frac{e^{\beta U(\mathbf{w}_j)}p(\mathbf{w}_j,t)-e^{\beta U(\mathbf{w}_i)}p(\mathbf{w}_i,t)}{||\mathbf{w}_j-\mathbf{w}_i||} \quad , \tag{5.14}$$

where V_i is the volume of cell I_i and A_{ij} is the interfacial area between I_i and I_j. In order to be consistent in the use of extensive quantities in the model, we rewrite the probability density p in terms of total probability in a cell P, such that

$$p(\mathbf{w}_i,t) = P_i(t)/V_i \quad . \tag{5.15}$$

Substituting the above into Eq. (5.14) gives

$$\dot{P}_i(t) = \sum_{j\neq i} A_{ij}\frac{D(\mathbf{w}_i)+D(\mathbf{w}_j)}{2}\; e^{-\beta[U(\mathbf{w}_i)+U(\mathbf{w}_j)]/2}\;\frac{e^{\beta U(\mathbf{w}_j)}P_j(t)/V_j-e^{\beta U(\mathbf{w}_i)}P_i(t)/V_i}{||\mathbf{w}_j-\mathbf{w}_i||} \tag{5.16a}$$

$$= \sum_{j\neq i}\left\{A_{ij}\frac{D(\mathbf{w}_i)+D(\mathbf{w}_j)}{2}\cdot\frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_i)-U(\mathbf{w}_j))\right]}{V_j||\mathbf{w}_j-\mathbf{w}_i||}P_j(t)\right\}- \\ \left\{\sum_{j\neq i} A_{ij}\frac{D(\mathbf{w}_i)+D(\mathbf{w}_j)}{2}\cdot\frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_j)-U(\mathbf{w}_i))\right]}{V_i||\mathbf{w}_j-\mathbf{w}_i||}\right\}P_i(t) \quad . \tag{5.16b}$$

This equation is key to our discretization scheme as it expresses the Smoluchowski equation through the values of D and U at the centers of the Voronoi cells as well as through the cell volumes, areas of the connecting faces and center-center distances between cells. The equation obeys detailed balance such that it ensures the existence of an equilibrium state given by the Boltzmann distribution. The equation is of great value as it provides the simplest possible account for the geometry of the Voronoi cells in the context of a discretized diffusion model reproducing the Smoluchowski equation in the limit of vanishingly small cells.

Finally, we define the coefficents of $P_j(t)$ and $P_i(t)$ to be R_{ij} and R_{ii} respectively, such that Eq. (5.16b) reads

$$\dot{P}_i(t) = \sum_{j \neq i} R_{ij} P_j(t) + R_{ii} P_i(t) \quad , \tag{5.17}$$

which can be written as a linear kinetic equation

$$\dot{\mathbf{P}}(t) = \mathbf{R} \cdot \mathbf{P}(t) \quad . \tag{5.18}$$

The rate matrix $\mathbf{R}$ arising in Eq. (5.18) has three important properties.
1. $R_{ij}P_j(t)$ is the rate of probability inflow from I_j to I_i and $R_{ii}P_i(t)$ is the rate of outflow from I_i to its nearest neighbors. By observing the terms in Eq. (5.16b), one will find that the total flow rate to other cells from I_i, given by $\sum_{i \neq j} R_{ji} P_i(t)$, is equal to $R_{ii}P_i(t)$, the outflow rate from I_i, thus ensuring particle conservation.
2. The solution of (5.18) relaxes to a stationary, i.e., equilibrium, distribution $\mathbf{P}^0$, which is given by the Boltzmann distribution

$$P_i^0 \equiv Z^{-1} V_i \, e^{-\beta U(\mathbf{w}_i)} \quad , \tag{5.19}$$

where Z is the partition function. By construction, the principle of detailed balance is obeyed:

$$\begin{aligned}
R_{ij} P_j^0 =& \left\{ A_{ij} \frac{D(\mathbf{w}_i) + D(\mathbf{w}_j)}{2} \cdot \frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_i) - U(\mathbf{w}_j))\right]}{V_j ||\mathbf{w}_j - \mathbf{w}_i||} \right\} Z^{-1} V_j \, e^{-\beta U(\mathbf{w}_j)} && \text{(5.20a)} \\
=& \frac{A_{ij}}{2Z} \left[D(\mathbf{w}_i) + D(\mathbf{w}_j) \right] \frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_i) + U(\mathbf{w}_j))\right]}{||\mathbf{w}_j - \mathbf{w}_i||} && \text{(5.20b)} \\
=& \left\{ A_{ji} \frac{D(\mathbf{w}_i) + D(\mathbf{w}_j)}{2} \cdot \frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_j) - U(\mathbf{w}_i))\right]}{V_i ||\mathbf{w}_j - \mathbf{w}_i||} \right\} Z^{-1} V_i \, e^{-\beta U(\mathbf{w}_i)} && \text{(5.20c)} \\
=& R_{ji} P_i^0 \quad . && \text{(5.20d)}
\end{aligned}$$

3. Reflective or absorptive boundary conditions can be imposed on any cell. If I_i is reflective, then the rates of inflow to and outflow from I_i are zero. Hence, we set

$$R_{ij} = R_{ji} = 0 \; \forall \; j = 1, 2, \ldots, N \; , \; j \neq i \quad . \tag{5.21}$$

If instead I_i is absorptive, then probability may flow in, but not out of I_i, whence

$$R_{ji} = 0 \; \forall \; j = 1, 2, \ldots, N \; , \; j \neq i \quad . \tag{5.22}$$

5.2.3 Solution of the Rate Equation

The next step is to solve Eq. (5.18), given the initial distribution $\mathbf{P}(0)$. The approach adopted here is to solve for the eigenvalues and eigenvectors of $\mathbf{R}$, which yield, together with the initial condition, an exact solution of Eq. (5.18). Thus, the only source of error due to time discretization is the assumption that a particle begins each time step being completely equilibrated within its current cell [131].

Depending on the number of cells N in $\mathcal{I}$, solving for the entire matrix $\mathbf{R}$ at once can be computationally expensive. We briefly discuss the complexity involved in Section 5.3. A better alternative is to solve for the diffusive behavior locally, as done in the framework of the Brownian dynamics algorithm [32, 157, 186], as well as in the MSM [131]. For this purpose, we make use of the fact that for a particle initially in I_i, the extent of diffusion is effectively limited to a region characterized by the radius $r_{\Delta t} = \sqrt{2nD\Delta t}$ about $\mathbf{w}_i$ in time step Δt, where n is the number of spatial dimensions, ignoring presently a possible drift of probability due to the non-zero local force $-\nabla U(\mathbf{x})$.

For each I_i, we consider the cell centers contained within some radius r_{restrict} of $\mathbf{w}_i$ and construct a restriction $\mathbf{R}^{(i)}$ of $\mathbf{R}$ to these cells. For our purposes, we set r_{restrict} to $2r_{\Delta t}$. $\mathbf{R}^{(i)}$ is comprised of only elements of $\mathbf{R}$ associated with probability transfers between cells within the $2r_{\Delta t}$ radius (see Fig. 5.1a for an illustration in the 2D case). For the purpose of the local computation, we index the matrix elements of $\mathbf{R}^{(i)}$ differently from those of $\mathbf{R}$ such that in general holds $R^{(i)}_{jk} \neq R_{jk}$. For the sake of bookkeeping, we henceforth use the local index notation lmn for local elements of $\mathbf{R}^{(i)}$ while we continue to employ ijk for the global indices. We also introduce the permutation $\sigma^{(i)}$ that maps the global index k of a cell to the corresponding local index n specific to the restricted region centered on cell I_i, namely

$$\sigma^{(i)}(k) = n \quad , \tag{5.23}$$

so that by construction holds

$$R^{(i)}_{\sigma^{(i)}(j)\,,\,\sigma^{(i)}(k)} = R_{jk} \quad . \tag{5.24}$$

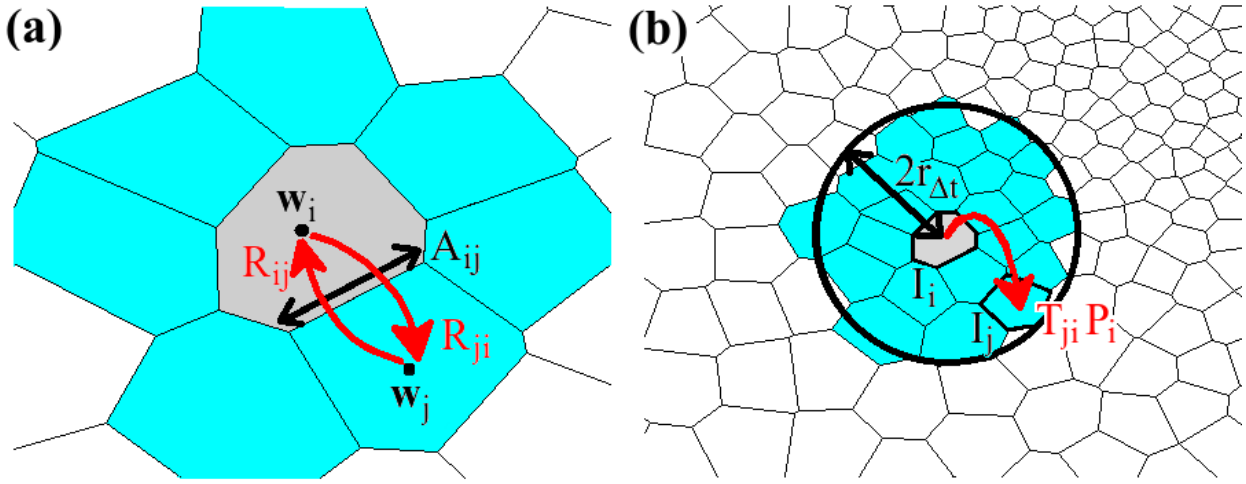

Figure 5.1: Discretization scheme for the case of two-dimensional diffusion. **(a)** Rate matrix components R_{ij} and R_{ji} give the rates of probability flow between cells I_i and its nearest neighbors, indexed by j, and are dependent on geometric properties such as the inter-cell distance $||\mathbf{w}_i - \mathbf{w}_j||$, the interfacial area A_{ij} and the cell volumes V_i and V_j (not shown in diagram). **(b)** Illustration of local probability flow from cell I_i to other cells (indexed by j) in its neighborhood (encircled area). In the deterministic approach, $T_{ji}P_i(t)$ gives the amount of probability that flows from cell I_i to cell I_j between times t and $t + \Delta t$. In the stochastic approach, T_{ji} gives the probability that a particle in cell I_i at time t hops to cell I_j between times t and $t + \Delta t$.

In order to preserve particle conservation, reflective boundary conditions are applied to the bordering cells of each restricted neighborhood, and $R_{ll}^{(i)}$'s are re-calculated so that the sum over elements in any column of $\mathbf{R}^{(i)}$ equals zero.

The problem to be solved then is

$$\dot{\mathbf{P}}^{(i)}(t) = \mathbf{R}^{(i)} \cdot \mathbf{P}^{(i)}(t) \quad , \tag{5.25}$$

where $\mathbf{P}^{(i)}(t)$ is the vector with components representing the probability in each locally indexed cell. Henceforth the superscript (i) will denote quantities that have been re-indexed locally. The `eig` Matlab subroutine is used to calculate the eigenvalues $\lambda_n^{(i)}$ and corresponding eigenvectors $\nu_n^{(i)}$ of $\mathbf{R}^{(i)}$, where $n = 1, 2, \ldots, \dim(\mathbf{R}^{(i)})$. The solution of Eq. (5.25) is then

$$\mathbf{P}^{(i)}(t) = \sum_n \alpha_n^{(i)} \exp(\lambda_n^{(i)} t) \nu_n^{(i)} \quad , \tag{5.26}$$

where $\{\alpha_n^{(i)}\}$ are scalar constants to be determined. The initial distribution defines the

coefficients $\alpha_n^{(i)}$ through

$$\mathbf{P}^{(i)}(0) = \sum_n \alpha_n^{(i)} \nu_n^{(i)} \quad . \tag{5.27}$$

In case that the particle is initially in cell i, holds

$$P_n^{(i)}(0) \; = \; \delta_{\sigma^{(i)}(i)\,,\,n} \quad . \tag{5.28}$$

Since $\mathbf{R}^{(i)}$ is not symmetric in general, $\{\nu_n^{(i)}\}$ is not expected to be an orthogonal set of vectors. However, $\mathbf{R}^{(i)}$ is similar to a symmetric matrix $\tilde{\mathbf{R}}^{(i)}$ under the transformation

$$\tilde{\mathbf{R}}^{(i)} = (\mathbf{M}^{(i)})^{-1} \mathbf{R}^{(i)} \mathbf{M}^{(i)} \quad , \tag{5.29}$$

where $\mathbf{M}^{(i)}$ is specified by $M_{lm}^{(i)} = \delta_{lm} [\, V_l^{(i)} \exp(-\beta U(\mathbf{w}_l^{(i)}))\,]^{1/2}$. Thence,

$$\begin{aligned}
\tilde{\mathbf{R}}_{lm}^{(i)} &= (M^{(i)})_{ll}^{-1} \, R_{lm}^{(i)} \, M_{mm}^{(i)} && \text{(5.30a)} \\
&= [\, V_l^{(i)} \exp(\beta U(\mathbf{w}_l^{(i)}))\,]^{1/2} && \text{(5.30b)} \\
&\quad \left\{ A_{lm}^{(i)} \frac{D(\mathbf{w}_l^{(i)}) + D(\mathbf{w}_m^{(i)})}{2} \cdot \frac{\exp\left[-\frac{\beta}{2}(U(\mathbf{w}_l^{(i)}) - U(\mathbf{w}_m^{(i)}))\right]}{V_m^{(i)} ||\mathbf{w}_m^{(i)} - \mathbf{w}_l^{(i)}||} \right\} && \text{(5.30c)} \\
&\quad [\, V_m^{(i)} \exp(-\beta U(\mathbf{w}_m^{(i)}))\,]^{1/2} && \text{(5.30d)} \\
&= A_{lm}^{(i)} \frac{D(\mathbf{w}_l^{(i)}) + D(\mathbf{w}_m^{(i)})}{2\sqrt{V_l^{(i)} V_m^{(i)}}\, ||\mathbf{w}_l^{(i)} - \mathbf{w}_m^{(i)}||} && \text{(5.30e)} \\
&= \tilde{\mathbf{R}}_{ml}^{(i)} \quad . && \text{(5.30f)}
\end{aligned}$$

Thus, $\tilde{\mathbf{R}}^{(i)}$ is symmetric, so that its eigenvectors $\{\tilde{\nu}_n^{(i)}\}$ are orthogonal. To find $\tilde{\nu}_n^{(i)}$, one observes that from the definition $\mathbf{R}^{(i)} \nu_n^{(i)} = \lambda_n^{(i)} \nu_n^{(i)}$ follows

$$\begin{aligned}
\tilde{\mathbf{R}}^{(i)} (\, (\mathbf{M}^{(i)}\,)^{-1} \nu_n^{(i)}) &= (\, (\mathbf{M}^{(i)})^{-1} \mathbf{R}^{(i)} \mathbf{M}^{(i)}\,)(\, (\mathbf{M}^{(i)})^{-1} \nu_n^{(i)}\,) && \text{(5.31a)} \\
&= (\mathbf{M}^{(i)})^{-1} \mathbf{R}^{(i)} \nu_n^{(i)} && \text{(5.31b)} \\
&= \lambda_n^{(i)} (\, (\mathbf{M}^{(i)})^{-1} \nu_n^{(i)}\,) \quad . && \text{(5.31c)}
\end{aligned}$$

Hence, $\tilde{\nu}_n^{(i)} = (\mathbf{M}^{(i)})^{-1} \nu_n^{(i)}$ is an eigenvector of $\tilde{\mathbf{R}}^{(i)}$ with eigenvalue $\lambda_n^{(i)}$. The orthogonality condition of the eigenvectors $\{\tilde{\nu}_n^{(i)}\}$ reads

$$\frac{\tilde{\nu}_l^{(i)} \cdot \tilde{\nu}_m^{(i)}}{|\tilde{\nu}_l^{(i)}| \cdot |\tilde{\nu}_m^{(i)}|} = \delta_{lm} \quad . \tag{5.32}$$

Putting Eq. (5.32) into Eq. (5.27) gives

$$\frac{(\,(\mathbf{M}^{(i)})^{-1}\mathbf{P}^{(i)}(0)\,)\cdot\tilde{\nu}_l^{(i)}}{|\tilde{\nu}_l^{(i)}|^2} = (\mathbf{M}^{(i)})^{-1}\sum_n \alpha_n^{(i)}\nu_n^{(i)}\cdot\frac{\tilde{\nu}_l^{(i)}}{|\tilde{\nu}_l^{(i)}|^2} \tag{5.33a}$$

$$= \sum_n \alpha_n^{(i)}\frac{\tilde{\nu}_n^{(i)}\cdot\tilde{\nu}_l^{(i)}}{|\tilde{\nu}_l^{(i)}|^2} \tag{5.33b}$$

$$= \sum_n \alpha_n^{(i)}\delta_{ln} \tag{5.33c}$$

$$= \alpha_l^{(i)} \quad . \tag{5.33d}$$

On condition that the particle is initially in cell I_i, we calculate the coefficients $\alpha_l^{(i)}$ by putting Eq. (5.28) into Eq. (5.33) and obtaining

$$\alpha_l^{(i)} = \sum_m\sum_n (M^{(i)})^{-1}_{mn}\,\delta_{\sigma^{(i)}(i)\,,n}\frac{(\tilde{\nu}_l^{(i)})_m}{|\tilde{\nu}_l^{(i)}|^2} \tag{5.34}$$

$$= \sum_m (M^{(i)})^{-1}_{mm}\,\delta_{\sigma^{(i)}(i)\,,m}\frac{(\tilde{\nu}_l^{(i)})_m}{|\tilde{\nu}_l^{(i)}|^2} \tag{5.35}$$

$$= \frac{(M^{(i)})^{-1}_{\sigma^{(i)}(i)\,,\,\sigma^{(i)}(i)}(\tilde{\nu}_l^{(i)})_{\sigma^{(i)}(i)}}{|\tilde{\nu}_l^{(i)}|^2} \quad . \tag{5.36}$$

Setting $t = \Delta t$, we thus obtain for the probability distribution in the neighborhood of the initial position after one time step

$$\mathbf{P}^{(i)}(\Delta t) = \sum_n \alpha_n^{(i)}\exp(\lambda_n^{(i)}\Delta t)\nu_n^{(i)} \quad . \tag{5.37}$$

Using the local transition probabilities given by the elements of $\mathbf{P}^{(i)}$ for every $i \in \{1, 2, \ldots, N\}$, we construct the transition matrix $\mathbf{T}$, where each element T_{ji} is the transition probability of a particle moving from cell I_i to cell I_j (see Fig. 5.1b), given by

$$T_{ji} = P^{(i)}_{\sigma^{(i)}(j)}(\Delta t) \quad . \tag{5.38}$$

The framework developed thus far can be used for both deterministic and stochastic simulations. In the deterministic case, one uses the transition matrix to propagate a probability distribution in time:

$$\mathbf{P}(t + \Delta t) = \mathbf{T}\cdot\mathbf{P}(t) \quad . \tag{5.39}$$

The deterministic approach was employed in the biological case involving two-dimensional diffusion of a nascent peptide chain within the SecY channel [185], as described in Section 5.5.

In the stochastic case, a random number generator iterates the position of each diffusing particle based on the relevant probabilities given by $\mathbf{T}$. At a given time t, suppose a given particle is in cell I_i. The position of the particle at the next time step $t+\Delta t$ is chosen, through the use of the `rand` random number generator `rand`, to be I_j with probability T_{ji}. Subsequent iterations of these algorithmic steps propagate the particle along its trajectory during the simulation. This method is closely related to the Brownian dynamics method [157, 186].

5.3 Computational Efficiency

The simulation process applied in our numerical solutions can be divided into three major phases according to computational expense: discretization of the system, calculation and solution of the rate matrix, and Brownian motion algorithm for the diffusive displacement of particles. In the discretization phase, the most expensive task to be performed is the adaptation of cell centers, which requires updating N positions in S adaptation steps. Hence, the complexity for the first phase goes as $O(N \cdot S)$.

In the second phase, the bottleneck occurs during the solution of the rate matrix, more specifically during the calculation of eigenvectors and eigenvalues for the neighborhood within radius $r_{\text{restrict}} = 2r_{\Delta t}$ of each cell center $\mathbf{w}_i$. Given the average density ρ, the approximate number of cells within each neighborhood is $\frac{4}{3}\pi(2r_{\Delta t})^3\rho = \frac{32}{3}\pi\rho(6D\Delta t)^{3/2}$. A modest ballpark estimate for the complexity of eigenvector expansion is $O(n^3)$. For N cells, this estimate gives $O(N\rho^3(D\Delta t)^{9/2})$ complexity for the eigenvector expansion phase. By this estimate, we justify the algorithmic step of solving restricted matrices as opposed to solving the entire rate matrix: depending on the characteristics of the system and grid chosen, it is often the case that the $O(N^3)$ complexity of solving the entire matrix overshadows that of solving the restricted matrices. In the final phase, given particle number M and a total simulation time t_{total}, the complexity is $O(M \cdot t_{total}/\Delta t)$.

The computationally most expensive phase is typically the second one. Fortunately, the calculations for each cell neighborhood can be performed independently of those for other neighborhoods, and, thus, are amenable to parallelization. Furthermore, eigenvector expansion algorithms themselves can sometimes be parallelized. Hence, there is much potential for the reduction of the overall computation time needed.

5.4 Simple Validation Test Cases

Three series of simulations of simple systems were performed for the purpose of validation. Each series corresponded to a different simulated system, and consisted of multiple simulation sets, with each set consisting of simulations in which a particular simulation parameter was varied. The analytic behavior for each system simulated is available for comparison with the simulation results.

In Series 1 of the validation trials, the model was tested for the bulk diffusive behavior of particles. Particles were initialized at the center of a spherically symmetric system of large enough radius that no particle reached the boundary throughout the duration of each trial. Each particle was allowed to freely diffuse (corresponding to a potential $U(\mathbf{x}) = 0$) as its displacement was recorded over time. The mean square displacement of all the particles was then calculated as a function of time. The mean square displacement of a 3D diffusing particle from its initial position ($t = 0$) to its position at time t is $< \Delta x^2(t) >= 6Dt$.

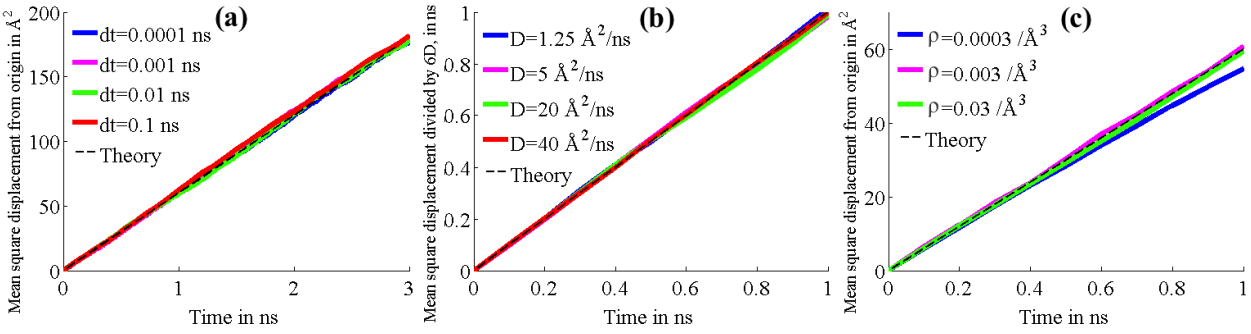

Figure 5.2: Results for Series 1 tests. Except for $dt = 0.0003/\text{Å}$ in (c), graphs in each test set are almost identical and obscure each other as a result. **(a)** Mean square displacement for varying time steps. **(b)** Mean square displacement for varying diffusion coefficients. **(c)** Mean square displacement for varying grid densities.

In Set 1A, each trial had a sample size of 10^4 particles. The diffusion coefficient was set to $D = 10$ Å^2/ns, the grid resolution was $\rho_1 = 0.03$ /Å^3 within radius 25 Å of the center, scaled linearly from 25 Å to 35 Å down to $\rho_2 = 0.003$ /Å^3 at which the density remains fixed up to the reflective system boundary at 60 Å. The time step was varied across values $dt = 0.0001, 0.001, 0.01, 0.1$ ns. As seen in Fig. 5.2(a), the simulation agrees well with theory regardless of the time step used. A 10^4-particle sample size was also used for Set 1B, in which case the grid resolution was as in Set 1A, $\rho_1 = 0.03$ /Å^3 and $\rho_2 = 0.003$ /Å^3, the time step was set to $dt = 0.1$ ns, while the diffusion coefficient was varied across values $D = 1.25, 5, 10, 40$ Å^2/ns. The results presented in Fig. 5.2(b) show that the simulation is

also accurate across different values of the diffusion coefficient. In Set 1C, the parameters used were a sample size of 10^4, $D = 10$ Å^2/ns, $dt = 0.1$ ns, $\rho_2 = 0.0003$ /Å^3 while the variable parameter was $\rho_1 = 0.0003, 0.003, 0.03$ /Å^3. The results shown in Fig. 5.2(c) show that the simulated behavior of bulk diffusion was robust over a wide range of grid resolutions, breaking down only at low grid densities on the order of 10^{-4} /Å^3 or less.

In Series 2, absorptive and reflective boundary conditions were imposed: cells within the spherical shell of radius $r_1 = 10$ Å were set to be absorptive while cells outside the shell of radius $r_2 = 30$ Å were set to be reflective. Figure 5.3(c) shows the geometry of the system. Particles were initialized in cells with centers within 0.5 Å of the shell of radius $r_i = 20$ Å, and then allowed to diffuse until all particles had left the system. The particle count was tracked as a function of time and compared against the theoretical behavior, given by Eq. (5.65c) derived in Section 5.4.1, and found in Carslaw and Jaeger [187] (second edition, Eqs. (12-15) on p. 367).

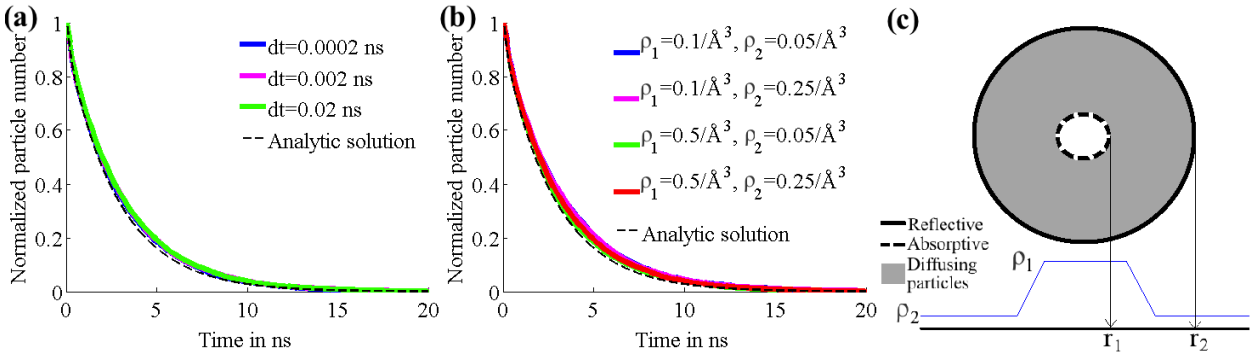

Figure 5.3: Results for Series 2 tests. Graphs in each test set are almost identical and obscure each other as a result. **(a)** Particle number for varying time steps. **(b)** Particle number for varying grid densities. **(c)** (top) Schematic of system used for Series 2; (bottom) radial profile of grid density used to model the system.

In Set 2A, the sample size was 10^4 for each trial, the diffusion coefficient $D = 150$ Å^2/ns, the grid density $\rho_1 = 0.5$ /Å^3 within radius 15 Å of the center and $\rho_2 = 0.25$ /Å^3 from 20 Å to the system boundary at 35 Å, with the value scaling down linearly between 15 Å and 20 Å; the time step was varied across values $dt = 0.0002, 0.002, 0.02$ ns. The results shown in Fig. 5.3 closely approximate the results from the analytical solutions with differences in time step size resulting in negligible differences in results on the time evolution of particle number. In Set 2B, the sample size was 10^4, $D = 150$ Å^2/ns, $dt = 0.02$ ns, and grid density was varied across values $\{\rho_1, \rho_2\} = \{0.1$ /Å$^3, 0.05$ /Å$^3\}$, $\{0.1$ /Å$^3, 0.25$ /Å$^3\}$, $\{0.5$ /Å$^3, 0.05$ /Å$^3\}$, $\{0.5$ /Å$^3, 0.25$ /Å$^3\}$.

In a final series of trials, the system and parameters used in Sets 3A and 3B were the same as those in Sets 2A and 2B. However, a linear potential $U(\mathbf{x}) = \alpha||\mathbf{x}||$ was imposed, where $\alpha = 0.2\ k_{\mathrm{B}}T/\text{Å}$. The results for the respective sets are shown in Fig. 5.4 and compared with the results from the analytic solution, given by Eq. (5.66) in Section 5.4.2. Again, the numerical results are accurate over the range of values of parameters tested.

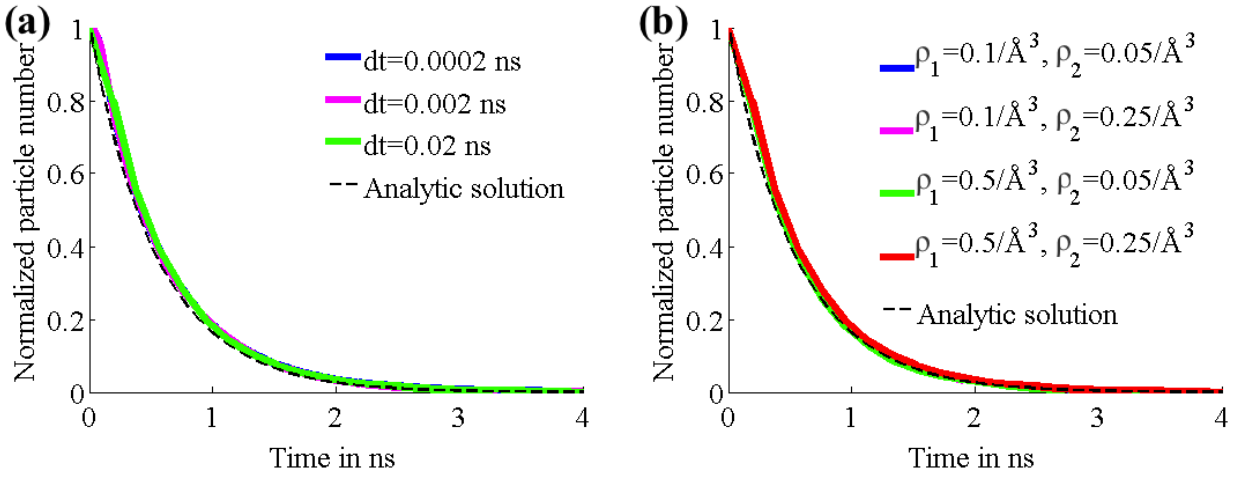

Figure 5.4: Results for Series 3 tests. Graphs in each test set are almost identical and obscure each other as a result. **(a)** Particle number for varying time steps. **(b)** Particle number for varying grid densities.

The time step-independence of the results in Figs. 5.2-5.4 is not unexpected, since the solution (5.37) of the restricted matrix is exact. In theory, the accuracy of the MSM description decreases with a decrease in the size of the time step [131] because of the implicit assumption that the equilibration time of the particle within the cell at the beginning of each time step is negligible. The implication for our model is that accuracy is independent of time step size above a certain minimum time step size, subject to other parameter values, in particular the radius of the restricted region r_{restrict}, which must be set large enough that the particle's movement is not significantly hindered by its boundaries. The other source of inaccuracy is error due to discretization as one assumes a uniform distribution in the grid cells. This error may be detected by varying either the rate at which particles move from cell to cell or by varying the grid resolution. In this regard, the results of Set 1C suggest that boundary interactions are a more significant source of discretization error than is bulk diffusion.

5.4.1 Analytical solution for validation Sets 2A and 2B

Here we provide the solution of the diffusion problem for test set 2. The solution is available from Carslaw and Jaeger [187] (second edition, Eqs. (12-15) on page 367), but rather than explaining the terms and constants in the complex solution expression stated by the authors, we derive the solution here as an optimal means of communication to the reader not versed in diffusion theory.

The system is spherically symmetric, with particles diffusing in the space between two concentric spheres of radii r_1 and r_2, as shown in Fig. 5.3(c). At time $t = 0$, particles are uniformly distributed on a spherical surface of radius r_i, with $r_1 < r_i < r_2$. The diffusion coefficient D is taken to be constant.

The free diffusion equation is given by

$$(\nabla^2 - D^{-1}\partial_t)p(\mathbf{r}, t) = 0 \quad . \tag{5.40}$$

Using the ansatz

$$p(\mathbf{r}, t) = e^{\omega t} Y_l^m(\theta, \phi) R(r) \quad , \tag{5.41}$$

one obtains for the radial dependence

$$\left[\partial_r^2 + \frac{2}{r}\partial_r - \frac{l(l+1)}{r^2} - \frac{\omega}{D}\right] R(r) = 0 \quad . \tag{5.42}$$

The above equation is known as the spherical Bessel equation, the solutions of which are of the form

$$p(\mathbf{r}, t) = \sum_{k,m,l} e^{\omega t} Y_l^m(\theta, \phi)(A_{kl} j_l(kr) + B_{kl} n_l(kr)) \quad , \tag{5.43}$$

where j_l and n_l are the spherical Bessel functions, k is defined through $k^2 = \frac{-\omega}{D}$, $Y_l^m(\theta, \phi)$ represents a set of functions called the spherical harmonics, and A_{kl} and B_{kl} are constants.

Due to spherical symmetry, only the $l = 0$ term contributes to the solution, so that

$$p(\mathbf{r}, t) = \sum_k e^{\omega(k) t} R_k(r) \quad , \tag{5.44}$$

where

$$R_k(r) = A_k j_0(kr) + B_k n_0(kr) \quad , \tag{5.45}$$

$$j_0(kr) = \frac{\sin(kr)}{kr} \quad , \tag{5.46}$$

$$n_0(kr) = -\frac{\cos(kr)}{kr} \quad , \tag{5.47}$$

and A_k, B_k are constants to be determined.

In our description we assume Dirichlet and Neumann boundary conditions at $r = r_1$ and $r = r_2$, respectively. At $r = r_1$ holds

$$R_k(r_1) = 0 \quad . \tag{5.48}$$

For the convenience of calculation, a constant phase may be introduced without loss of generality:

$$\sin(kr) \to \sin[k(r - r_1)] \quad , \tag{5.49}$$

$$\cos(kr) \to \cos[k(r - r_1)] \quad . \tag{5.50}$$

Then, Eq. (5.48) implies that $B_k = 0 \; \forall \; k \in \mathcal{R}$. Thus it holds

$$R_k(r) = A_k \frac{\sin[k(r - r_1)]}{kr} \quad . \tag{5.51}$$

At $r = r_2$ the Neumann boundary condition is assumed

$$\partial_r R_k(r)|_{r=r_2} = 0 \quad . \tag{5.52}$$

This condition reads

$$-\frac{\sin[k(r_2 - r_1)]}{kr_2^2} + \frac{\cos[k(r_2 - r_1)]}{r_2} = 0 \quad , \tag{5.53}$$

or

$$\tan[k(r_2 - r_1)] = kr_2 \quad . \tag{5.54}$$

The numerical solution of Eq. (5.54) gives a countably infinite set of values of k. For the sake

of clear notation, let n index these values. Accordingly, we re-index the following terms:

$$R_k(r) \to R_n(r) \quad , \tag{5.55}$$

$$\omega(k) \to \omega_n \quad , \tag{5.56}$$

$$A_k \to A_n \quad . \tag{5.57}$$

Defining the inner product in solution space as

$$< R_{n_1} | R_{n_2} > = \int_{r_1}^{r_2} r^2 R_{n_1}(r) R_{n_2}(r) dr \quad , \tag{5.58}$$

it can be verified that $\{R_n \,|\, n \in \mathbb{Z}\}$ is an orthogonal set. The normalization factor is obtained by evaluating

$$A_n^{-2} < R_n | R_n > = (k_n)^{-2} \int_{r_1}^{r_2} \sin^2[k_n(r - r_1)] dr \tag{5.59a}$$

$$= \frac{1}{2k^2} \left(-r_1 + \frac{k_n^2 r_2^3}{1 + k_n^2 r_2^2} \right) \quad . \tag{5.59b}$$

Hence, the inner product can be explicitly written

$$< R_{n_1} | R_n > = A_n^2 \frac{\delta_{n_1,n}}{2k^2} \left(-r_1 + \frac{k_n^2 r_2^3}{1 + k_n^2 r_2^2} \right) \quad . \tag{5.60}$$

The complete solution is then

$$p(\mathbf{r}, t) = \sum_n e^{\omega_n t} R_n(r) \quad , \tag{5.61}$$

where $\omega_n = -k_n^2 D$.

In the case of the assumed initial condition $p_0(\mathbf{r}) = (4\pi r_i^2)^{-1} \delta(r - r_i)$ follows

$$\sum_n R_n(r) = (4\pi r_i^2)^{-1} \delta(r - r_i) \quad . \tag{5.62}$$

Taking the inner product gives

$$< R_m | R_m > = (4\pi r_i^2)^{-1} \int_{r_1}^{r_2} r^2 R_m(r) \delta(r - r_i) dr \quad , \tag{5.63}$$

from which follows

$$A_m = A_m(4\pi < R_m|R_m >)^{-1} R_m(r_i) \tag{5.64a}$$

$$= \frac{k_m^2}{2\pi}\left(-r_1 + \frac{k_m^2 r_2^3}{1+k_m^2 r_2^2}\right)^{-1} \frac{\sin[k_m(r_i - r_1)]}{k_m r_i} \quad . \tag{5.64b}$$

Finally, the expression of the number of surviving particles as a function of time is

$$\Sigma(t) = \int_{r_1}^{r_2} dr\, 4\pi r^2 p(r,t) \tag{5.65a}$$

$$= 4\pi \int_{r_1}^{r_2} dr \sum_n A_n r e^{-k_n^2 Dt} \frac{\sin[k_n(r - r_1)]}{k_n} \tag{5.65b}$$

$$= 4\pi \sum_n e^{-k_n^2 Dt} \frac{A_n r_1}{k_n^2} \quad . \tag{5.65c}$$

5.4.2 Numerical mean first passage time description for validation sets 3A and 3B

In the case of a linear potential $U(r) = \alpha r$, a closed-form analytic solution for the time-dependent diffusion probability does not exist. Instead, the time dependence of the probability distribution was captured for the test purpose by the mean first passage time approximation [140]. In this approximation, the normalized surviving particle count is given by

$$\Sigma(t) = e^{-t/\tau(r_i)} \quad , \tag{5.66}$$

where $\tau(r_i)$ is the mean first passage time given by the expression

$$\tau(r_i) = \int_{r_1}^{r_i} dr\, (Dr^2)^{-1} e^{\alpha r} \int_r^{r_2} dr'\, r'^2 e^{-\alpha r'} \quad . \tag{5.67}$$

$\tau(r_i)$ is evaluated through numerical integration.

The mean passage time description through Eqs. (5.67,5.66) is known to be a good approximation to the decay of the total probability of still unreacted particles [140] and, hence, Eqs. (5.67,5.66) can serve as a test of the numerical scheme suggested in the present study.

5.5 Biological Application 1: Membrane Insertion of Nascent Peptide Chains through SecY Translocon

Membrane proteins are hardly ever inserted directly into the lipid bilayer. During translation, a nascent peptide chain enters the translocon channel, SecY, which either provides passage to the other side of the bilayer or inserts the chain into the bilayer itself.

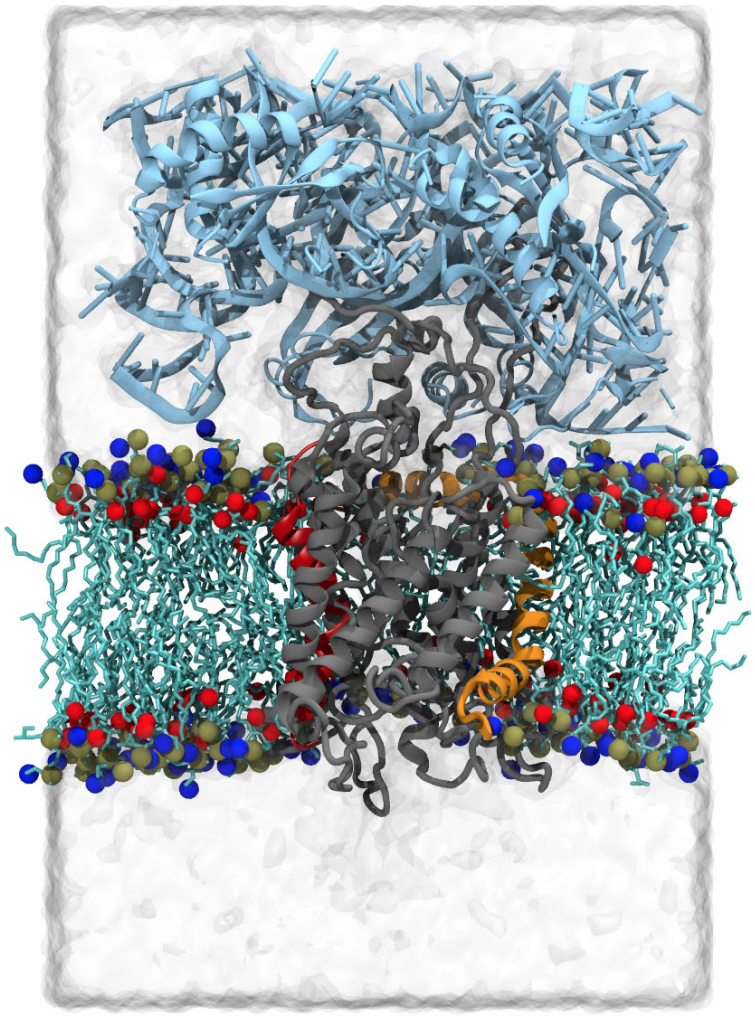

Figure 5.5: Simulated system used for MD simulations. SecY is shown in grey, SecE in orange, and the retained portion of the ribosome in light blue. The lipid tails are in cyan with selected phosphorus, nitrogen, and oxygen atoms of the head groups displayed as brown, blue, and red spheres, respectively. The water box for the periodic system is in light grey.

There has been much speculation over the mechanism that determines which path a given peptide sequence will take. It is known that the process of translocation between the in-

terior of SecY and the membrane is controlled in significant part by the hydrophobicity of the residues in the chain, with the apparent free energy of transfer between translocon and the membrane characterized by a so-called "biological hydrophobicity scale" [188, 189]. However, the free energy for all 20 amino acids fall in a narrow range of around -1 to 3 kcal/mol, as compared to a scale based on partitioning directly between water and hydrophobic solvent[190, 191], which ranged from -5 to 15 kcal/mol [192]. The relative order of the amino acids are largely the same in both scales.

The difference in apparent free energy range between the two scales has been attributed to the different molecular environment that each scale corresponds to [193]. However, the molecular states corresponding to the biological hydrophobicity scale are not known, making this hypothesis difficult to confirm. Another view is that kinetic factors come into play, specifically, that translocation in the SecY case is a non-equilibrium process and that the measured free energies do not reflect the actual thermodynamic equilibrium [194, 195].

The present study proposes a model that characterizes the membrane insertion process, based on both energetic and kinetic considerations, that is consistent with experimental data obtained so far. PMF calculations along the path from the center of the channel to the exterior of the channel gate were used in conjunction with a 2D diffusion simulation to describe the dynamics of the nascent chain in the membrane plane. From the simulations, a two-state scheme was formulated, whereby the nascent chain transitions from being inside the channel to being in the membrane with a probability corresponding to the apparent free energy of translocation. Furthermore, as a result of interplay between energetic and kinetic factors, the compression of the apparent free energy of translocation relative to the free energy difference between the two states in the potential of mean force was observed.

5.5.1 Molecular Dynamics Simulations

For construction of the simulation model, the starting structure was taken from Frauenfeld et al. [71] (PDB: 3J00/3J01), consisting of SecYE bound to a ribosome (see Fig. 5.5), together with a nascent chain which includes the signal anchor (SA). For computational economy, portions of the ribosome 20 Å away from SecY were removed, since the region of interest is in the membrane around SecY itself. Atoms around the loose ends of the truncated ribosome were harmonically restrained during simulation. Next, nascent chain residues that were not part of the SA were removed. The channel was then embedded in a 75%/25% POPE/POPG membrane containing 200 lipids. Altogether, the system consisted of around 120,000 atoms.

Equilibration of the system was performed with NAMD [137], using the CHARMM22 force field with CMAP corrections [56] for proteins and CHARMM36 force field for lipids [95]. The temperature was 325 K. The integration time step was 2 fs, while short- and long-range non-bonded interactions (separated by a cutoff of 12 Å) were evaluated every 1 and 3 time steps, respectively. The particle-mesh Ewald method was employed for calculating long-range electrostatics. Equilibration was performed at NPT with a pressure of 1 atm. Subsequent MD simulations are performed at NVT and different temperatures (when stated) in some cases, but otherwise employ the same parameters as the equilibration.

Long-time simulations on Anton proceed from the equilibrated structure obtained via NAMD. Force field and time-stepping procedure were the same as in the NAMD simulations. Simulations were run in NVT using the Berendsen coupling scheme, with long-range electrostatics calculations handled by the k-Gaussian Split Ewald method on a $64 \times 64 \times 64$ grid. The cutoff was determined independently for each simulation, but typically was around 13 Å. The total time of all Anton simulations was about 30 μs.

MD simulations of the models were used to investigate several hypotheses on the mechanism of translocation. In summary, the MD simulations support the view that membrane insertion of a helix from within SecY is dependent on hydrophobicity of the helix, but this dependence is mediated by lipid contact by the helix rather than by hydrophobicity-induced opening of the gate as previously suggested [196]. The simulations also showed that gate opening is instead correlated with binding of a ribosome to SecY, and that the qualitative behavior of helices of various hydrophobicities is consistent with a two-state thermodynamic model of the system. Crucial to appreciating the findings of the present study, these simulations and their results are described in detail in Section 5.5.2.

The constructed model was also used for umbrella sampling calculations to quantitatively describe the energetics of membrane insertion of the helix from SecY. The umbrella sampling calculations were performed for the SA, polyLeu, and polyGln helices. These calculations utilized the colvars module of NAMD [197], with 26 windows spaced about 1 Å apart, beginning at the center of SecY and ending in the membrane. For each helix, the total simulation time was 250 ns, so that 750 ns was spent altogether on umbrella sampling. PMFs were extracted from the resulting histograms using WHAM [198], and are shown in Fig. 5.6B. As expected, the hydrophobic helices SA and polyLeu favor insertion into the membrane by 1-2 and 4-5 kcal/mol, respectively, while the polyGln helix favors remaining in the channel by more than 10 kcal/mol. Furthermore, the magnitudes of the free energies correspond to the hydrophobicity order of the helices. It should be noted that these free

energies likely do not represent the apparent free energies of membrane insertion, ΔG_{app}, since the latter is measured experimentally in a more complex environment where the nascent chain is longer and attached to the ribosome, and translation and translocation occur in tandem with membrane insertion.

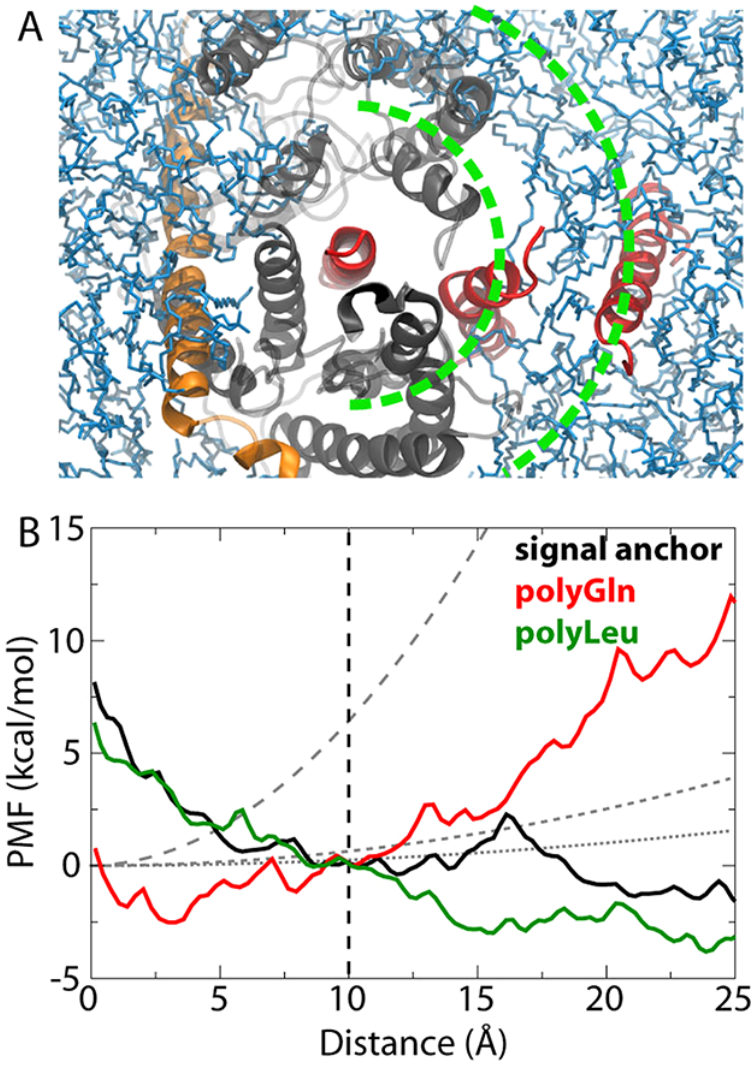

Figure 5.6: Potential of mean force for helix exit from SecY into the membrane. **(A)** SecY is shown from the cytoplasmic side in gray and orange with the membrane in blue. A substrate helix is shown in red at different positions along its exit, although only one helix was present at any given time. The green dotted lines are at $r = 12$ Å and $r = 25$ Å. **(B)** PMFs for the SA (black), polyLeu (green), and polyGln (red) helices as a function of distance from the channel center. The gray dashed lines show, in order of decreasing dash size, the restraining potential used in the diffusion calculations at times t = 1 s, 10 s, and 25 s.

5.5.2 Molecular Dynamics Simulation Results

A set of MD simulations of the translocon system was performed to test the hypothesis that the opening and closing of SecY's lateral gate, which provides entry to the membrane, is controlled by the hydrophobicity of the nascent chain [196]. Thus, one would expect hydrophobic chains to open the gate and hydrophilic ones to close it. For this purpose, the system was simulated with nascent chains of different hydrophobicities in SecY, as well as with none. The chains chosen were the SA, polyLeu, polySer, and polyGln. Each chain was simulated for different initial states of the lateral gate, namely closed, partially open, and fully open. The state of the SecY gate was characterized by the distance between C_α atoms of Ser87 on TM2b and Phe286 on TM7 (see Fig. 5.7); the closed, partially open and fully open states corresponded to distances of 7-10 Å, 14 Å, and 27 Å, respectively. The simulations were run for durations between 0.5 to 2 μs.

Fluctuations in the gate opening was monitored during the simulations and plotted in Fig. 5.8. For all three initial gate opening states, no correlation was observed between the gate distance and the hydrophobicity of the nascent chain. In the same simulations, contact between the nascent chain and lipids in the membrane was also measured. Here, correlation was observed, with the hydrophobic chains (the SA and polyLeu) increasing interaction area with lipids and the hydrophilic ones (polySer and polyGln) decreasing the area (see Fig. 5.9). It was further observed that contact with lipids was not mediated through opening of the lateral gate, but rather by incursion of lipids into the channel (see Fig. 5.10). The results suggest that lipid interactions play a bigger role than gate opening and closing in deciding which peptide chains are inserted in to the membrane.

It has also been hypothesized that the binding of the ribosome to SecY predisposes the gate towards being open [199]. This suggestion is supported by numerous experimental findings, including structures that show partially-open SecY channels bound to different substrates [200, 201, 202] and electrophysiology measurements indicating persistent permeability of the ribosome-SecY complex to ions and small molecules after the nascent chain was removed [203, 204, 205]. 10-ns simulations have also demonstrated a slight bias towards opening when the closed SecY structure was bound to a ribosome [206]. Here, simulations were run with the ribosome present and absent, with both initially open and closed gates. For the closed gate simulations, ribosome binding was found to slightly increase gate separation (see Fig. 5.8D). For the open gate simulations, the unbound SecY was observed to start closing. These findings support the role that ribosome-binding plays in determining biasing the gate towards an open state.

In the case of a thermodynamic 2-state model, it is predicted that energetics would favor a spontaneous transition of a transmembrane helix from the SecY interior to the membrane [207]. This view is supported by the Frauenfeld et al. crystal structure showing the SA in proximity to the gate [71]. To further probe this model, 2.5 μs simulations of the ribosome-bound SecY containing the SA, polyLeu, polyGln, and the isolated KvAP S4 transmembrane segment, localized to the crystal structure position, were run. The latter was included because it is just above the threshold for membrane insertion [208]. Simulations were run at an elevated temperature of 353 K to accelerate movements into or out of SecY. As in previous simulations, the hydrophobic helices, the SA and polyLeu, were drawn by lipid interactions 4-5 Å out of SecY. It was also observed that the lateral gate closed behind the outgoing helix and the pore ring of SecY shrunk, thus preventing re-entry of the helix into the channel. In contrast, the S4 and polyGln helices pulled away from the gate, moving 5-7 Å towards the interior of the channel, which is predominantly hydrophilic [209, 91]. The behavior of the nascent chains in these simulations qualitatively support the thermodynamic 2-state model.

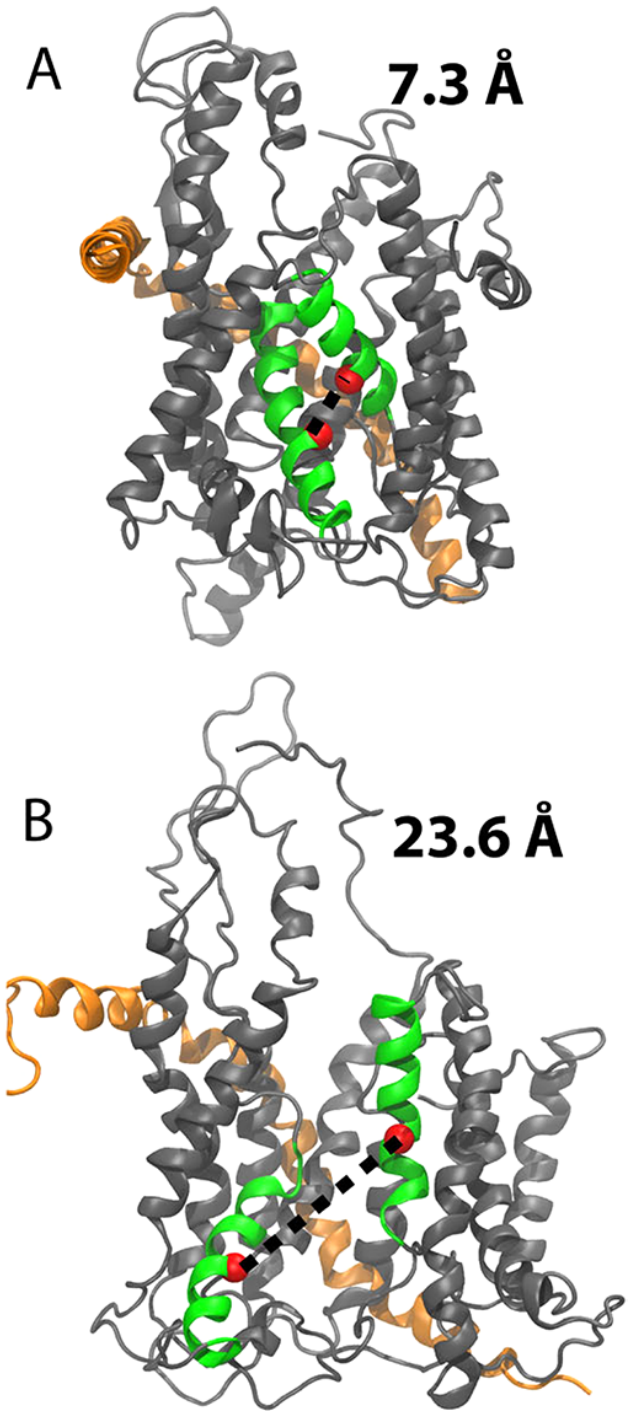

Figure 5.7: Lateral gate opening: SecYE shown in gray (SecY) and orange (SecE), with lateral gate helices TM2b and TM7 highlighted in green and residues Ser87 and Phe286 shown as red spheres. **(A)** Closed state of the gate (Ser87-Phe286 distance of 7.3 Å) [210]. **(B)** Open state from a membrane-protein-insertion intermediate structure [71].

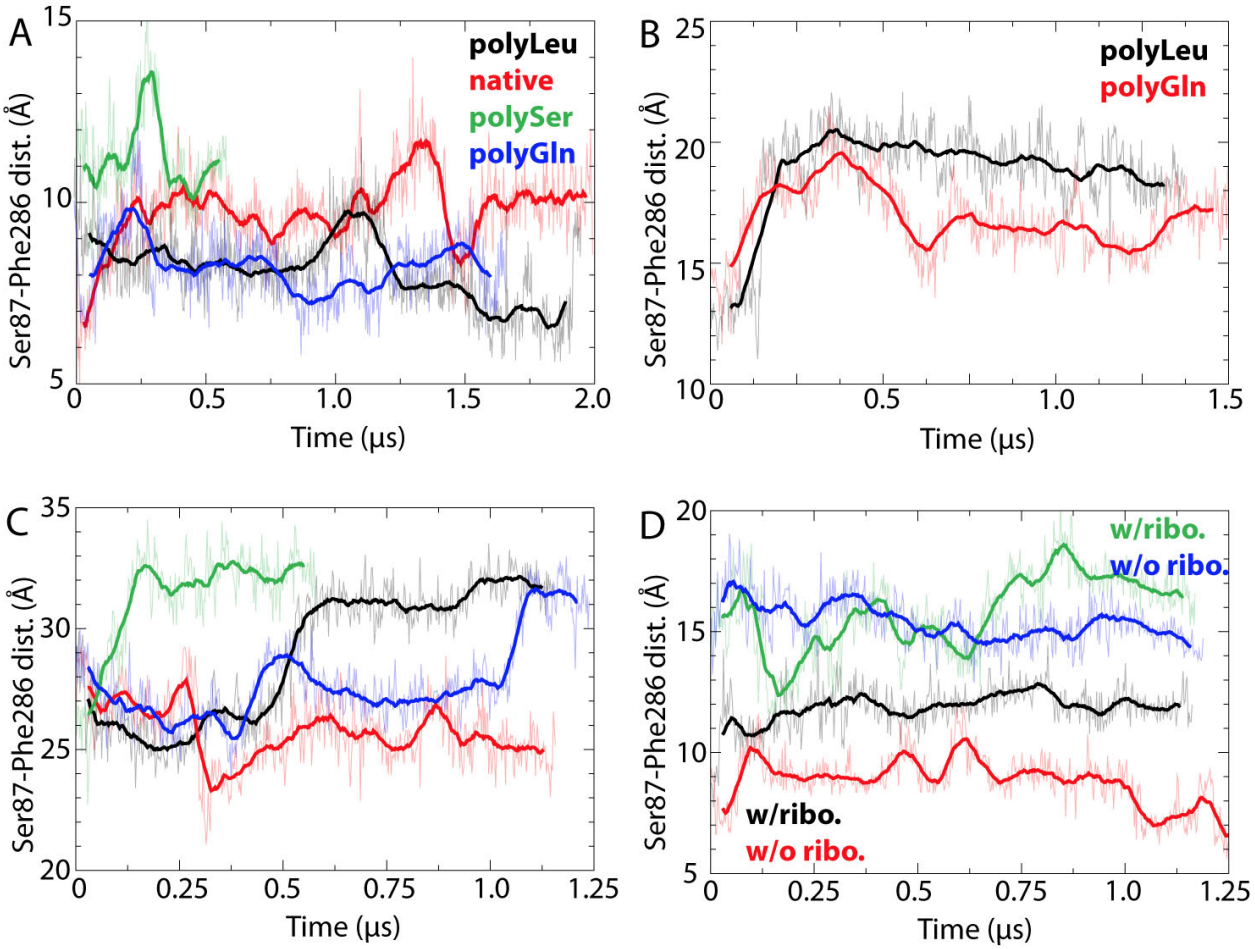

Figure 5.8: Plots of Ser87-Phe286 distance over time with different nascent helical TM segments embedded in SecY's central pore. **(A)** Initially closed SecY. Gate opening for hydrophobic helices (polyLeu and SA) are shown in black and red, respectively, with hydrophilic ones (polySer and polyGln) in green and blue. **(B)** Initial intermediate opening with polyLeu (black) and polyGln (red) inside. **(C)** Initially open, colored the same as in (A). **(D)** Empty SecY started in a closed state (red/black) and in a state of intermediate gate opening (blue/green), with and without the ribosome bound.

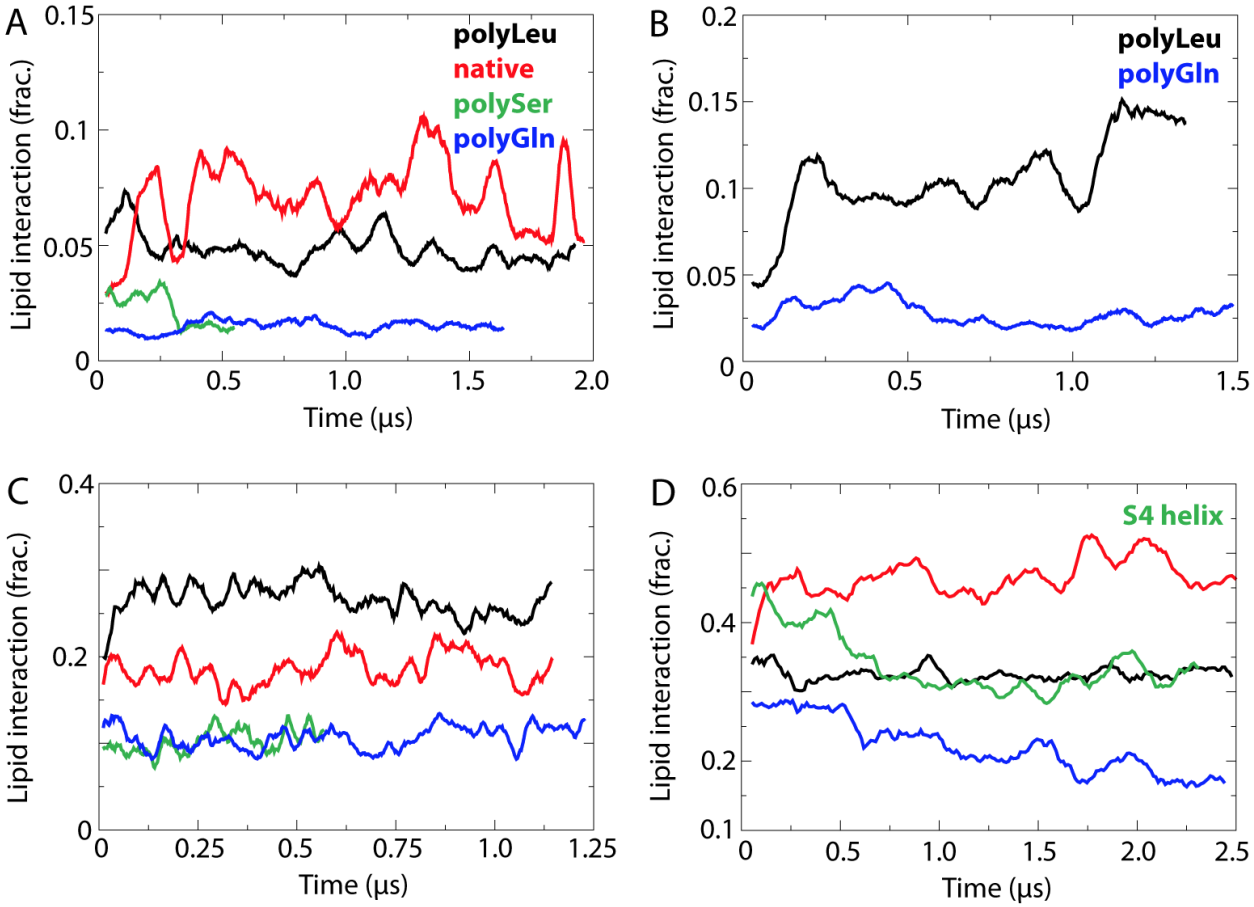

Figure 5.9: Interaction between lipids and a tested helix as a fraction of its total surface area. Lipid-interaction area for helices embedded in the center of the channel, corresponding to **(A)** Initially closed SecY. **(B)** SecY with an intermediate opening initially. **(C)** Initially open SecY. **(D)** Change in interaction area for helices initially positioned near the lateral gate. Helices are colored as in (A), except for polySer, which is replaced by the S4 helix here.

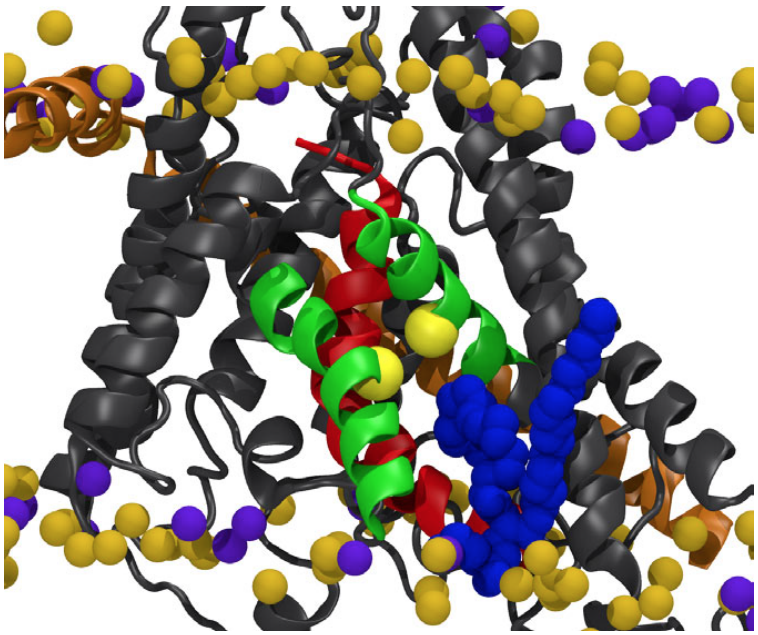

Figure 5.10: Incursion of a lipid into the closed channel after ∼0.5 μs. SecY is shown in grey, SecE in orange, and the lateral gate helices in green. The nascent SA helix (red) is still within the predominantly closed channel, indicated by the proximity of residues Ser87 and Phe286 (yellow spheres). A lipid contacting the SA across the otherwise closed gate is shown in a blue space-filling representation.

5.5.3 Diffusion Simulations

One kinetic factor that may contribute to the apparent free energy is the tethering of the nascent chain to the ribosome as it is being translocated. More specifically, the diffusion range of the nascent chain in the plane of the membrane is limited by the length of the chain between the ribosome and the SecY channel. Such a scenario could correspond, for example, to the case where the presence of a stop-transfer sequence in SecY halts translocation, allowing the nascent chain to accumulate outside SecY as translation proceeds [211].

A 2D deterministic model of diffusion in the plane of the membrane was constructed to explore the effect of tethering on membrane insertion probability. For this purpose, 2D Voronoi grid representing a circular membrane patch of radius 200 Å was constructed, with a region of constant high grid density within 20 Å of the origin, representing the interior of SecY (see Fig. 5.11), where a high level of detail is required to describe the geometry of the environment. The density falls linearly with distance from the center of the channel, at radii greater than 20 Å, so that the grid becomes coarse further away from SecY. An

occupancy map of SecY atoms was generated from the all-atom model, and a cross-section of the map was taken and overlaid on the grid. Grid cells that were occupied by SecY atoms were designated as having reflective boundary conditions applied to their surfaces, so as to create excluded volumes corresponding to SecY's geometry.

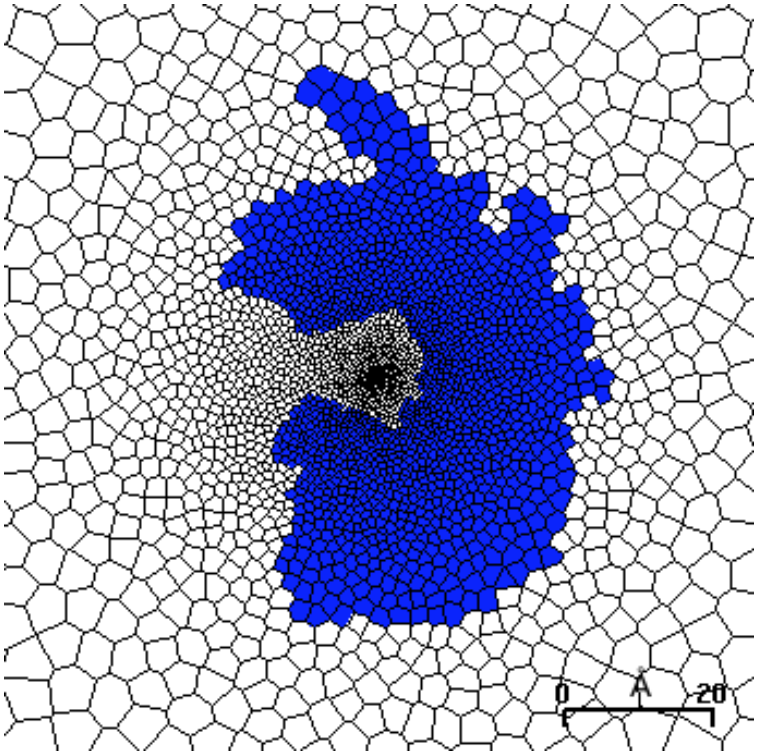

Figure 5.11: Voronoi tessellation of membrane plane segment, enlarged to show SecY cross-section (in blue). The density of cells is higher within the channel so as to describe diffusive behavior inside the channel in greater detail.

The helix is represented as a time-evolving probability density on the grid, under the influence of a 2D PMF obtained by extrapolating the 1D PMFs, obtained either by umbrella sampling or a linear approximation, radially about the center of the channel and adding a time dependent hard-shell potential that limits diffusion to within a circular region about SecY's center but gradually expands, mimicking the effect of extension of the nascent helix during translation. The ribosome-to-SecY portion of the helix is modelled as a freely jointed chain, so that the diffusion range is given by the average end-to-end distance, proportional to the square root of the number of residues between the ribosome and SecY. The rate of range expansion is then the rate of translation, which varies between 0.5 to 20 residues/s [212, 213, 211].

In production runs, the spatial distribution of the helix was initially confined to a cell approximately located at the center of SecY. The distribution is propagated in time, over a duration of typically 50 s. The helix was simulated by calculating the Boltzmann distribution over the diffusion space at each time-step. This Boltzmann model is justified by a validation test demonstrating with the deterministic kinetic diffusion model that the probability distribution of the helix reaches equilibrium within one time step (see Section 5.5.4). The Boltzmann model was favored over the kinetic diffusion model during production runs for computational expediency.

5.5.4 Algorithm Validation

The diffusion coefficient of each helix was assumed to be 250 $\text{Å}^2/\mu$s, as roughly estimated from the umbrella sampling simulations, and the time step used was 2 s. The first set of validation cases tested the ability of the deterministic kinetic diffusion model to reproduce diffusive behavior predicted by theory. Specifically, a probability distribution, initially confined to a cell near the origin, was propagated over time and the probability fraction falling outside a given radius R was tracked. Simulation parameters used were identical to those of production runs, except that a flat potential was used and the excluded region representing SecY was absent. The theoretically expected behavior can thus be calculated by solving the Einstein free diffusion equation. The results for various values of R, shown in Fig. 5.12, indicate agreement between the kinetic diffusion simulation and theory.

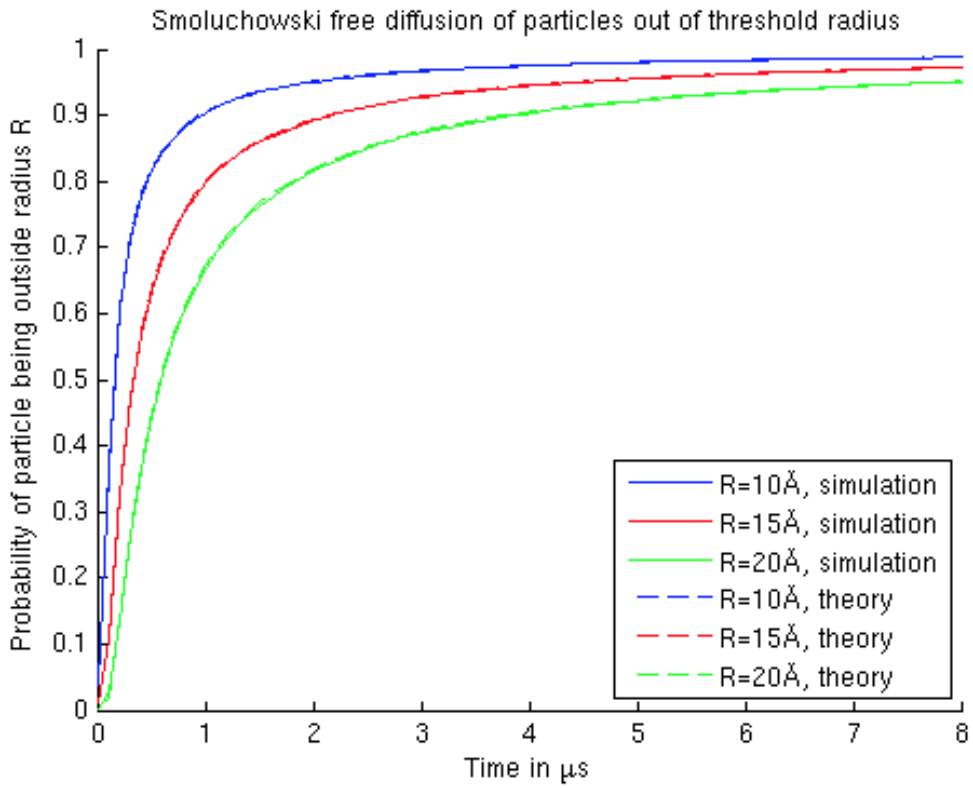

Figure 5.12: Results for free diffusion in two dimensions with system radius 200 nm; 5000 Voronoi cells in the distribution with a density of $\rho(r \mid r < 20 \text{ Å}) = \text{constant}$, $\rho(r \mid r > 20 \text{ Å}) \propto (r - 15 \text{ Å})^{-1}$, and diffusion coefficient 250 $\text{Å}^2/\mu\text{s}$. Proportion of particles outside radius r_{cutoff} were calculated for $r_{\text{cutoff}} = 10$ Å, 15 Å, and 20 Å.

The next validation test compared the outputs of the kinetic diffusion model and the Boltzmann model. Like the production runs, a linear radial PMF and a gradually expanding hard-shell limit on the diffusion range were used, in addition to the excluded region representing the SecY channel. The probability fraction outside a cutoff radius was monitored over time. The results, displayed in Fig. 5.13, show that the Boltzmann model closely tracks the kinetic diffusion model, indicating that the time step considered was sufficiently large for the system to be completely equilibrated each time the hard-shell limit was expanded. Owing to the computational ease of computing the Boltzmann distribution at each time step, the Boltzmann model was almost 12 times as fast as the kinetic diffusion model. Thus, the use of the Boltzmann model for production runs is justified.

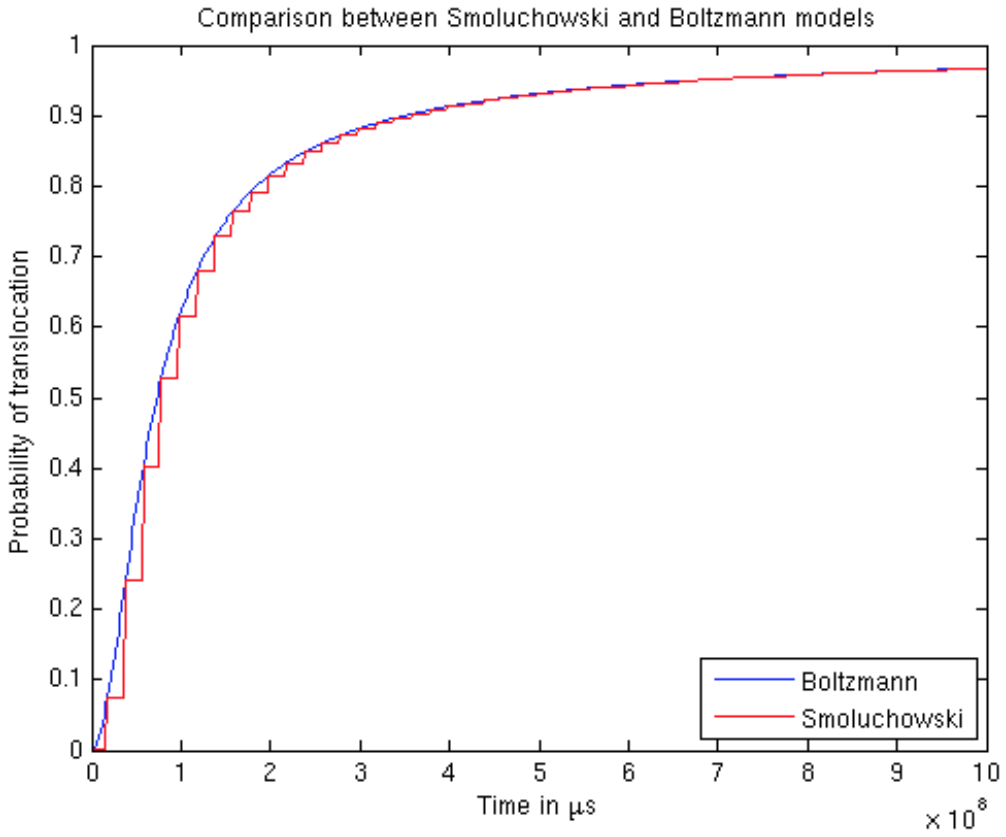

Figure 5.13: Results of Smoluchowski and Boltzmann simulations in a system of radius 2000 Å centered on the pore axis of the SecY channel. The parameters used include a time step 2 s, a PMF update time step 20 s for the Smoluchowski model, and a diffusion coefficient of 250 $\text{Å}^2/\mu\text{s}$. The potential of mean force used is $U(r,t) = (0.1\text{kcal}\,\text{mol}^{-1}\text{Å}^{-1})\,r + (0.0619\text{kcal}\,\text{s}\,\text{mol}^{-1}\text{Å}^{-2})\,r^2\,/\,t$. The graph displays the proportion of particles found outside radius $r_{\text{cutoff}} = 15$ Å.

Finally, it is expected that below a certain local grid resolution threshold, the geometry of the system would no longer be accurately described and simulation results would begin to diverge. To ensure that a suitable grid resolution is used in the production runs, two sets of test simulations were performed using the Boltzmann model. In the first set, the number of grid cells was kept constant, but the grid distribution function was varied. Here, it was found that the grid distribution function used in the production runs produced the same behavior as the other density functions tested (see Fig. 5.14A). In the next set of simulations, the grid distribution was kept constant, but the total number of grid cells was varied. It was found that simulation results began to diverge when the total number of cells fell below 1000 (see Fig. 5.14B), which is significantly smaller than the 5000 used in production runs.

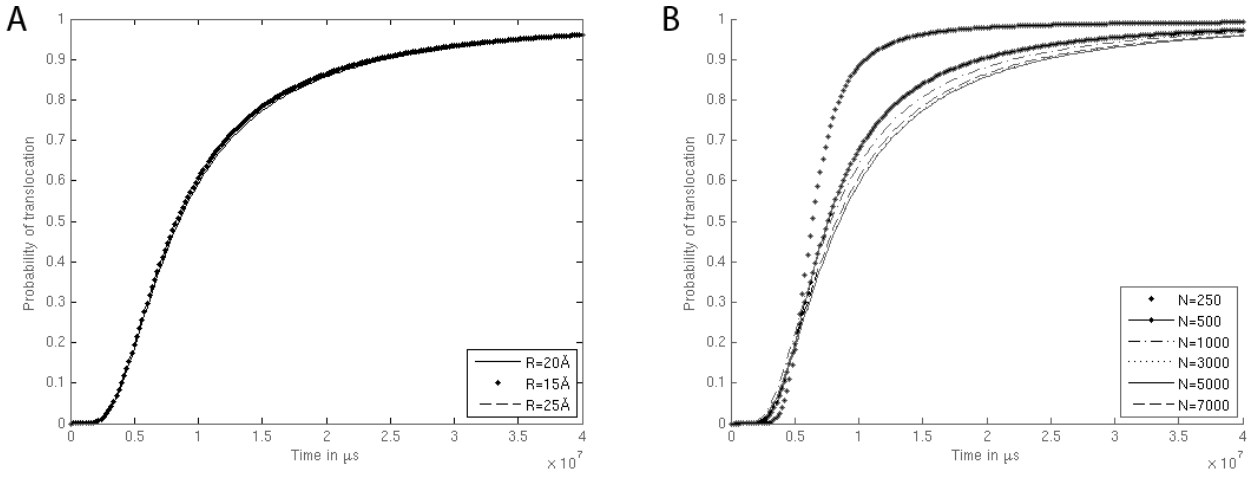

Figure 5.14: **(A)** Comparison of results using grid densities of the form $\rho(r \,|\, r < R) = \rho_0, \rho(r \,|\, r > R) \propto (r^3/4\,R)^{-1}$ for R=15, 20, 25 Å while keeping ρ_0 constant. This comparison tests for sensitivity to grid density gradients. **(B)** Comparison of results using R=20 Å and varying numbers of cells N. This comparison determines the threshold resolution beyond which results begin to diverge.

5.5.5 Results and Discussion

The first set of production runs investigated the effect of varying different parameters over typical values as determined by experiment. For each case, the instantaneous insertion probability was calculated as the fraction of the spatial probability distribution of the helix that lay outside a given cutoff radius from the center of SecY. Except for cases testing the effect of varying the cutoff, this radius is set to 15 Å to coincide with the gate opening. The results are shown in Fig. 5.15. In particular, note the strong influence of translation rate on the instantaneous insertion probability in the typical range of rates from 0.5 to 20 residues/s [212, 213, 211]. A slower rate of translation would hold the helix close to the center of SecY for a longer period of time, thereby slowing down membrane insertion. However, experiments found that the overall insertion probability was unchanged across the same range of translation rates [213], which can happen if, for example, if the translocation rate were coupled to the translation rate. Decreasing the translocation rate, which is equivalent to increasing the commitment time , increases the membrane-insertion probability, in agreement with experiment [214].

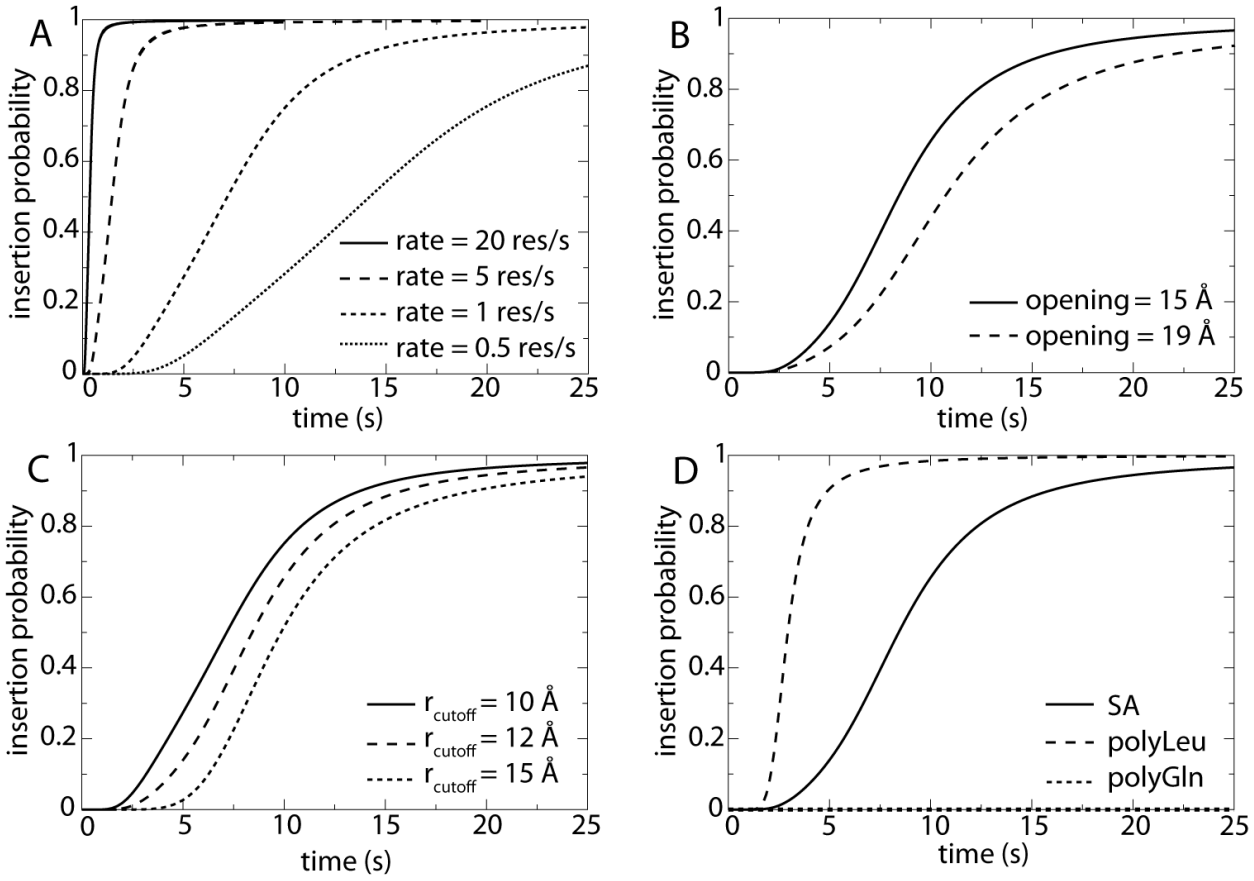

Figure 5.15: Probability of membrane insertion. The baseline parameters used are a translation rate of 1 res/s, a gate opening of 15 Å, and r_{cutoff} =12 Å. The results reflect insertion of SA as a function of **(A)** translation rate, **(B)** gate opening, and **(C)** rcutoff =12 . **(D)** Insertion of SA compared to the polyLeu and polyGln helices. The ranges of parameter values tested cover typical values as determined empirically. Note that size of the SecY gate opening in B and threshold location in C have small effects on the insertion probability, which is more strongly affected by rate of polypeptide chain lengthening and hydrophobicity of the polypeptide helix.

The next set of production runs investigated the effect of hydrophobicity on insertion probability. For this purpose, a set of 5 idealized linear PMFs (inset of Fig. 5.16) were employed. The slopes of the idealized PMFs were approximated from the realistic cases of the SA, polyLeu, and polyGln, but cover a wider hydrophobicity range. The effect of the smoothness of the idealized PMFs on the insertion probability is not expected to be significant. Note that the probability curves obtained from the first set of runs (Fig. 5.15) were smooth despite the noisiness of the PMFs of the SA, polyLeu and polyGln in Fig. 5.6B. The resulting time-dependent insertion probability curves are shown in Fig. 5.16A.

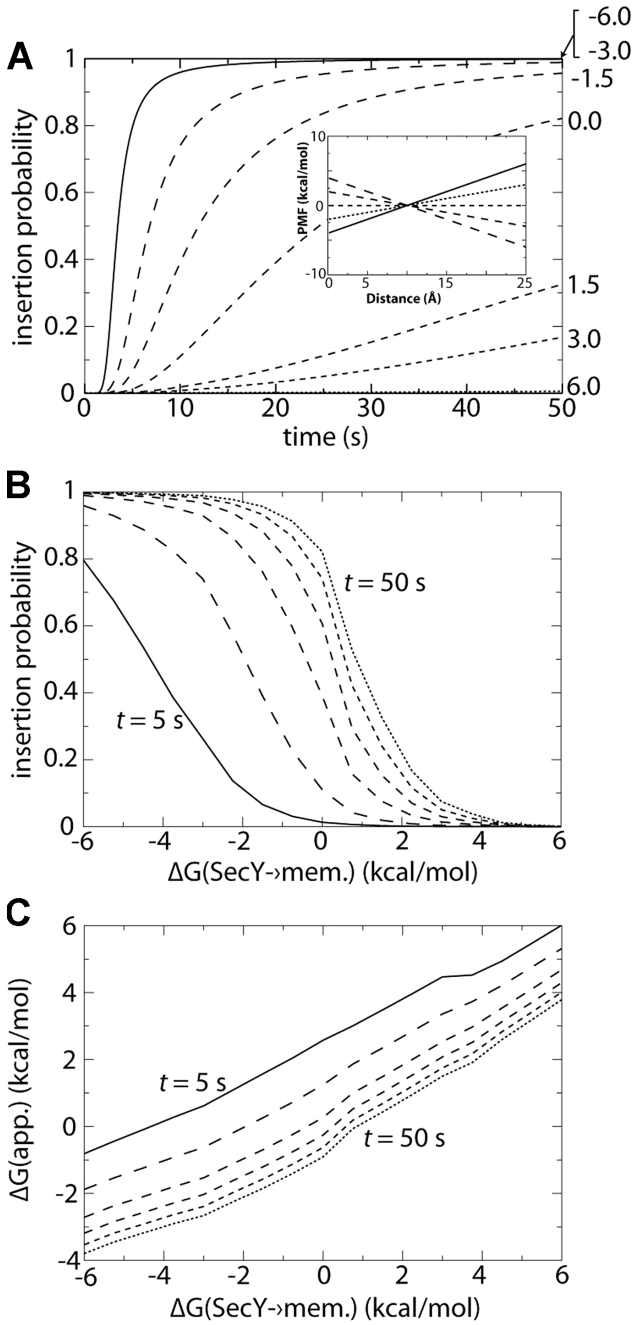

Figure 5.16: Membrane-insertion probability based on simplified PMFs. **(A)** Insertion probability as a function of time is plotted for linear PMFs of varying slope, shown in the inset plot. The corresponding $\Delta G(\text{SecY} \rightarrow \text{mem.})$ values using a reference point 15 Å into the membrane are given to the right of each curve. **(B)** Insertion probability as a function of $\Delta G(\text{SecY} \rightarrow \text{mem.})$ for commitment times, from left to right, of t = 5, 10, 20, 30, 40, and 50 s. **(C)** Relationship between ΔG_{app} and $\Delta G(\text{SecY} \rightarrow \text{mem.})$ for the same commitment times as in part (B).

The interplay of kinetics due to tethering and membrane insertion dynamics was elucidated through the matching of each of the resulting probability curve in Fig. 5.16A to a calculated free energy of insertion, denoted $\Delta G(\text{SecY} \rightarrow \text{mem.})$ and defined as the difference in PMF between the radii of 5 and 15 Å. Interestingly, the range of insertion probabilities is broadest around $\Delta G(\text{SecY} \rightarrow \text{mem.}) = 0$; in other words, the difference in insertion probability between, e.g., -1.5 and 1.5 kcal/mol is much greater than that between 3 and 6 kcal/mol. This enhanced range explains the observed sensitivity of marginally hydrophobic helices to a myriad of factors. For example, slowing translocation through the channel enhances membrane integration for mildly hydrophobic TM segments [214].

By reading along each value of time in Fig. 5.16A, a plot of insertion probability against $\Delta G(\text{SecY} \rightarrow \text{mem.})$ can be constructed. Plots for 6 values of time - 5, 10, 20, 30, 40, 50 s are shown in Fig. 5.16B. These curves should be interpreted as the dependence of insertion probability on the commitment time t, i.e. the time available to the helix to insert into the membrane before it is ejected out to the periplasm. Except for the 5-s curve, all curves are sigmoidal, similar to experimental insertion probabilities from which the biological hydrophobicity scale was determined [188]. In addition, the curves show asymptotic behavior towards $t = 50$ s, corresponding to the synthesis of 50 residues at the assumed 1 residue/s translation rate in the model. This number matches experiments demonstrating that stop-transfer efficiency plateaus at lengths greater than 40-50 residues [213].

Finally, for each curve in Fig. 5.16B, an apparent insertion free energy can be calculated by the definition

$$\Delta G_{\text{app}} = -k_B T \ln[p_{\text{ins}}(t)/p_{\text{sec}}(t)]\,, \tag{5.68}$$

where k_B is the Boltzmann constant, T is temperature, and p_{ins} and p_{sec} are the probabilities of insertion and secretion (into the periplasm) at time t, respectively. ΔG_{app} is plotted as a function of $\Delta G(\text{SecY} \rightarrow \text{mem.})$ in Fig. 5.16C. Note that the relationship is roughly linear in each case, with a slope of about 0.65. Hence, the apparent insertion-free-energy scale is compressed with respect to the SecY-to-membrane transfer free energy, which is itself already compressed with respect to the water-to-membrane transfer free energy. Hence, there are multiple factors contributing to the compression of the biological hydrophobicity scale with respect to other scales [192]. Furthermore, note that increasing the commitment time does not alter the slope of a line, but instead shifts its intercept to a lower value, thus decreasing the threshold for membrane insertion, as also observed experimentally [214].

Taken together, these simulations suggest that membrane insertion of polypeptide chains through SecY is governed not by thermodynamics alone, but in conjunction with kinetic

factors like the lengthening of the tethered chain during translation. It should be noted that these results are not intended to be interpreted as a complete representation of the real biological system, but to simply suggest qualitative trends in the membrane insertion probabilities. Among the biological features not accounted for in the diffusion model are movement of the helix from the channel into the periplasm, backsliding, retention of the helix near the channel due to interactions with from other channel substrate proteins, and opening and closing of the gate.

5.6 Biological Application 2: Ion Diffusion Through the Mechanosensitive Channel of Small Conductance

Mechanosensitive channels of small conductance (MscS) are a class of membrane channels that are gated by membrane tension. One such channel in *Eschericia coli* (ecMscS) is part of the bacterium's coping mechanism against osmotic stress. The increased tension in the membrane during osmotic stress activates ecMscS, allowing the passive efflux of cytoplasmic solutes, thereby mitigating a potentially fatal buildup of osmotic pressure [158].

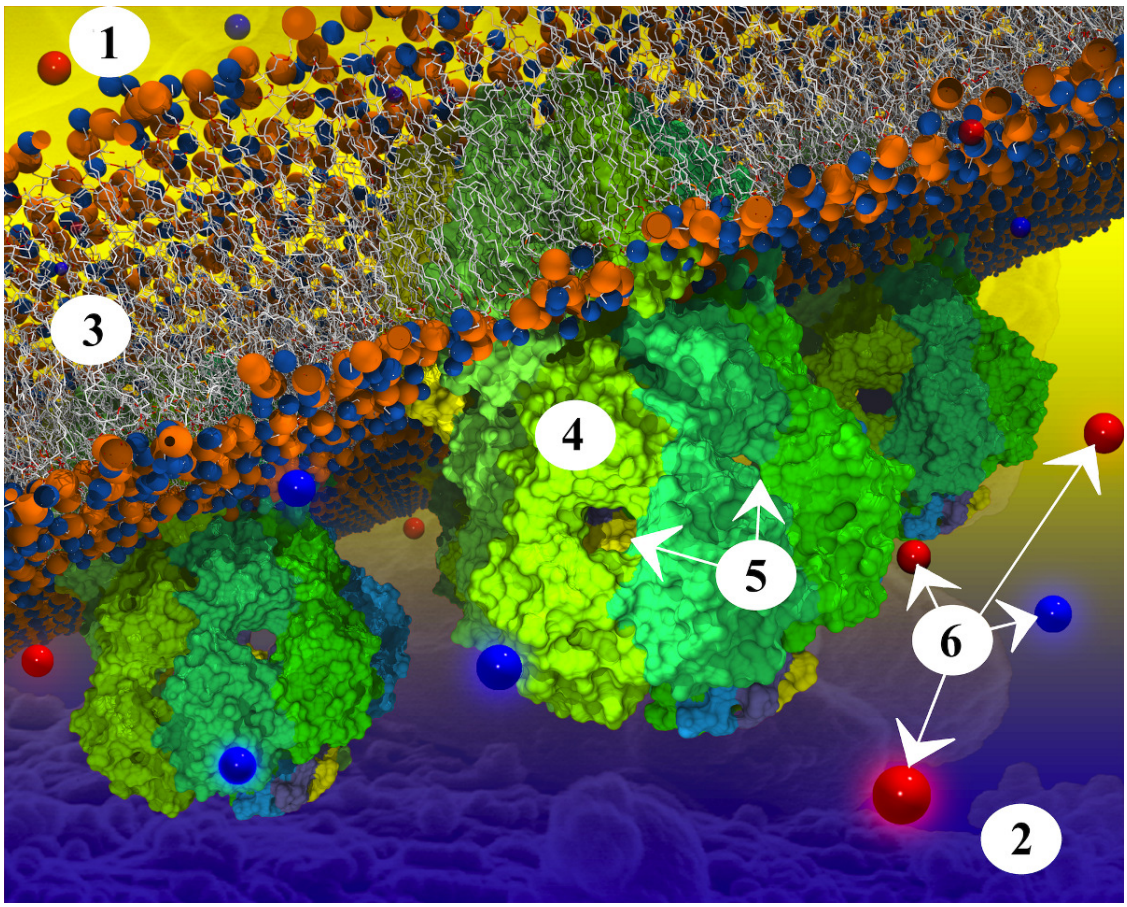

Figure 5.17: Three ecMscS channels embedded in a membrane. ecMscS is composed of seven identical subunits (each shown in a different color). ecMscS (shown in shades of green) has a transmembrane domain (faintly visible in the membrane in case of the ecMscS closest to the viewer), forming a channel that opens and closes in the presence or absence of significant osmotic pressure across the cell membrane. ecMscS also has a large extra-membrane domain pointing into the cell interior, the cytoplasm. This domain of ecMscS is called the cytoplasmic domain and is prominently visible in the figure for all three ecMscS proteins. The membrane is a bilayer of lipids; lipids are composed of head groups (shown as orange and blue spheres) and tails (shown as white lines). Ions (positive ions shown as blue spheres, negative ions as red spheres) diffuse through the bulk solvent, the cytoplasm, as seen here below the membrane. Ions enter the ecMscS cytoplasmic domains through the side openings into the domain interior and, in case of osmotic stress having induced an opening of the ecMscS trans-membrane channel, pass through the channel towards the space outside of the cell, the periplasm. The intricate geometry of the cytoplasmic domain, a roughly spherical interior connected to the cytosol through seven narrow openings, plays a determining role in the manner in which ions leave an osmotically challenged cell. The labels shown on the figure correspond to: 1. Periplasm; 2. cytoplasm; 3. membrane; 4. cytoplasmic domain of ecMscS; 5. side openings; 6. ions. Figure adapted from Ref. 3.

ecMscS, shown in Fig. 5.17, is a homoheptamer, consisting of a transmembrane domain and a cytoplasmic domain (CD). The CD is shaped like a balloon with seven lateral openings corresponding to its sevenfold symmetry. The lateral openings are just large enough for ions, the key osmolytes in *E. coli* cells, to pass through. The prevalent positive ions in *E. coli* cells are K^+ and Na^+, while the main negative ion is actually the amino acid glutamate, Glu^- [158]. An additional opening exists on the end of the CD distal to the membrane, but this opening is too narrow and hydrophobic for ions to pass through.

The function of the ecMscS CD is not well-understood despite its considerable size and large fraction of protein mass that goes into it. It is worth noting that CDs are a ubiquitous feature of ion channels, present in both Kv [215, 216, 217] and Kir [218, 219, 220] classes of potassium channels, some members of the family of voltage-gated sodium channels [221], the sodium-potassium pump [222], the ClC chloride channel [223], the nicotinic acetylcholine receptor and its homologs [224, 225], and the family of mechanosensitive channels [226, 227]. In most of these channels, the role of the CD has only in recent years become the subject of many investigative efforts [228, 229, 230, 231, 232, 233, 234, 158], as the importance of the respective CDs' roles in channel function become more evident. In particular, CDs are believed to play a role in regulating the diffusion of ions through the pores of certain channels.

In the case of ecMscS, one postulate is that the CD plays a role in gating and in this case would necessarily undergo a large conformational change between the open and closed states of the channel [235, 236, 237]. It has also been suggested that the CD stabilizes the structure of the channel [238, 239]. Another postulate is that the CD acts as a molecular filter that minimizes the loss of Glu^- solutes [227]. Such a filter may also encourage the efflux of cations and anions in pairs such that the efflux is electrically neutral, so as to maintain the cellular membrane potential. More recent studies compared ecMscS with channels with a similar CD structure, namely bacterial cyclic nucleotide-gated (bCNG) channels, MscS-Like proteins of *Arabidopsis thaliana* (MSL10) and *Thermoanaerobacter tengcongensis* MscS (TtMscS). bCNG channels display slight or no mechanosensitive gating response [240], suggesting that the CD plays at most a limited role in channel gating; the fact that MSL10 [241] and TtMscS [242] are highly anion-selective casts doubt on the view that the CD enforces current neutrality.

Previous attempts to simulate MscS function using MD and Monte Carlo methods showed high selectivity for anions [176]. These simulation results run counter to experimental measurements that indicate a much lower selectivity [243, 244, 245]. Barring inaccuracies associated with the simulation method, the results of experiment and simulation may be reconciled

if collective inter-ion interactions occur over time scales beyond the reach of the previous simulation methods used ($\sim 10\ \mu$s), that compensate for the channel's bias towards anions. Apart from such interactions, diffusive approach of the ions to the channel may also play an important role in the description of ion efflux, for example, if the time scale of the diffusive approach is larger than the time scale of passage through the channel. Simulated ion channel systems are typically not large enough to take into account the diffusive approach, which occurs over length scales of 10 - 100 nm. Addressing the needs for such long time and large length scales requires the simulation of ecMscS in a large box and necessitates the present method. We note that one can use a lower grid density to describe the large regions further away from the channel and a higher grid density for the channel and its vicinity.

However, since the proposed algorithm treats particles independently, the question of ion-ion interactions cannot yet be addressed. Rather, this simulation of ecMscS serves as a demonstration of the implementation of the kinetic diffusion algorithm, and to show that the results obtained are consistent with previous studies of the same system using the BioMOCA software and MD [176].

5.6.1 Setup of Molecular Dynamics Simulation

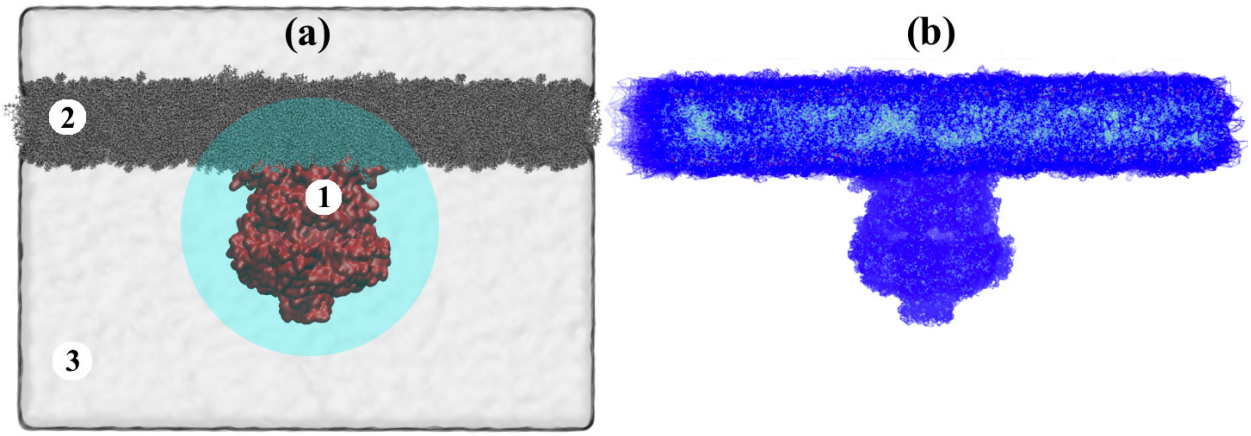

Figure 5.18: All-atom system used in the present study. **(a)** ecMscS (1) embedded in a membrane (2) and submerged in a waterbox (3). The highlighted circle represents a locus of 70 Å about the origin, in which a high grid density is used to model ion diffusion in detail near ecMscS. **(b)** Cells from the resulting grid that are associated with high PMF values and, hence, modelled as reflective barriers.

The system is described through an all-atom MD simulation with ecMscS embedded in the center of a 320 Å × 320 Å POPC membrane patch (see Fig. 5.18, immersed in a waterbox of dimension 316 Å × 317 Å × 230 Å. Ions are placed in the solvent in numbers according to physiological ion strengths [K^+], [Glu^-] and [Cl^-], such that the system is electrically neutral. The system is minimized and equilibrated in the presence of an electric field as described before [176] so as to widen the ecMscS pore relative to the opening seen in the crystal structure [227]. Full details of the all-atom system setup and simulation parameters are furnished in Section 5.6.2.

PMF maps of the system for both K^+ and Glu^- were extracted from a 240-ns equilibration run. Cross-sections of these maps are shown in Fig. 5.19. For the extraction the backbone of ecMscS was harmonically restrained. The distributions of K^+ and Glu^- ions were averaged over the entire course of the run. For the purpose of visualizing the PMF, regions where the distribution went to zero were assigned a minimal non-zero probability value, in order to prevent singularities from occurring when taking logarithms in the next step: the logarithm of the averaged distribution map, after normalization, gives the PMF map in units of k_BT. However, the assignment of non-zero probabilities was not used in the actual simulation, whereby grid cells corresponding to regions where the distribution is zero were simply excluded, i.e. no particles may transition into these cells.

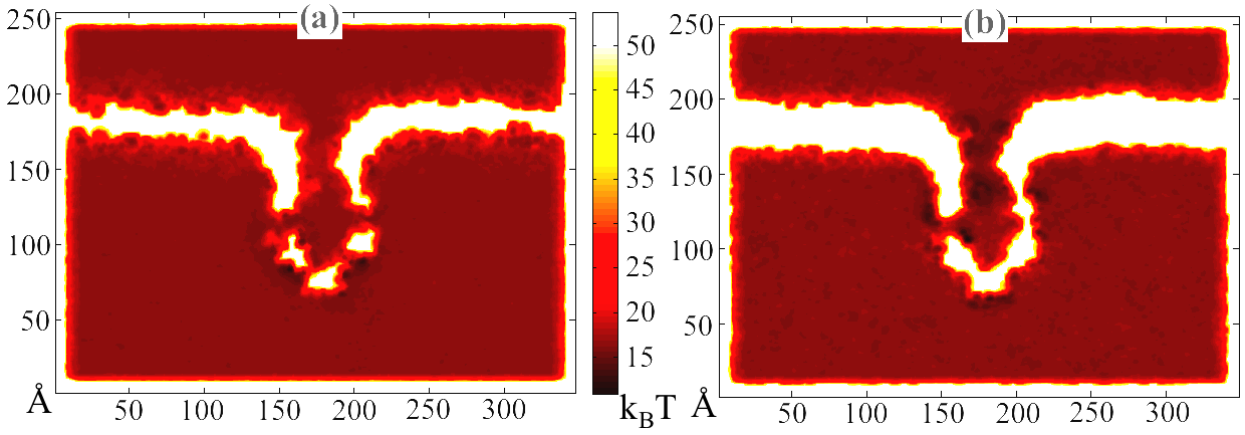

Figure 5.19: **(a)** Cross-section of potential of mean force maps for K^+ ions. **(b)** Cross-section of potential of mean force maps for Glu^- ions. The energy landscape for ions exhibit higher, more unfavorable values for positive ions inside the cytoplasmic domain and, conversely, lower, more favorable values for negative ions. Indeed, in the molecular dynamics simulations anions congregated in the cytoplasmic domain and sterically hindered the passage of cations.

The diffusion coefficients of K^+ and Glu^- were assumed to be constant in space. The

average value of each coefficient was obtained from trajectories arising from the MD simulation described above. The trajectories were divided into 0.02 ns intervals. In each interval, the mean square displacement of each ionic species between the beginning and the end of the interval was measured. The mean square displacement values were then averaged over all intervals. Thus, the diffusion coefficient D for each ionic species was calculated from the relation $< \Delta x^2(t) >= 6Dt$. The diffusion coefficients were found to assume the values $D_{\mathrm{K}} = 200$ Å^2/ns and $D_{\mathrm{Glu}} = 75$ Å^2/ns, which are in close agreement with the experimentally determined values [246, 247] of $D_{\mathrm{K}}^{\mathrm{expt}} = 196$ Å^2/ns and $D_{\mathrm{Glu}}^{\mathrm{expt}} = 75$ Å^2/ns (both measured in bulk solvent), respectively.

A grid representing the discretized system was built, with density $\rho_1 = 0.05$ /Å^3 within a radius of 70 Å (large enough to encapsulate the ecMscS, such that the center of the CD coincides with the center of the grid) and $\rho_2 = 0.01$ /Å^3 outside of a radius of 80 Å with the center of the CD being characterized through zero radius. The density was adapted linearly with the radius between 70 Å and 80 Å. The PMF map, obtained as values on a Cartesian grid, was cubic-interpolated to assign a PMF value to each cell center. Grid cells with the maximum PMF value, corresponding to regions where the ion distribution is zero, were designated as reflective boundaries. Consequently, these cells were excluded from the rate matrix calculation and solution, thus reducing the total computational cost; correspondingly, cells on the boundary of the system were assigned reflective surfaces. For a time step of $dt = 0.002$ ns, the rate and transition matrices were calculated, as outlined in Section 5.2.3.

5.6.2 Specifications for MD Simulation of ecMscS System

Model construction was performed using VMD [87]. The structure of ecMscS, solved through X-ray crystallography [227], was taken from PDB entry 1MXM. Asp, Glu, Lys and Arg were modeled as charged residues. The structure was then embedded in a 320 Å $\times$ 320 Å POPC membrane patch. Subsequently, protein-embedded membrane was immersed in a waterbox of dimension 316 Å $\times$ 317 Å $\times$ 230 Å and K^+ and Cl^- ions were placed in the solvent, such that the concentration of each ion was 200 mM, the physiological concentration of K^+ in the *E. coli* cytosol. The physiological concentration of Cl^- is on the order of 10 mM [248], while that of Glu^- has been reported to vary over a wide range of values [249]. Hence, enough Cl^- ions were kept in the system to give 10 mM concentration while the rest were switched with Glu^- zwitterions. Since ecMscS has a net charge of $+28e$, 28 K^+ ions were deleted to

neutralize the system. Due to Glu^- being larger in size than Cl^-, it was necessary to delete water molecules that overlapped with the Glu^- zwitterions. The final system contained 610,961 water molecules, 2997 lipids, 2307 K^+ ions, 100 Cl^- ions and 2235 Glu^- zwitterions. The total atom count of the system was 2,304,943.

MD simulations were run with NAMD 2.9 [137], using the TIP3P water model [139], and with the CHARMM36 [95] and CMAP-corrected CHARMM22 [56] force field for lipids and non-lipids, respectively. The time step was set to 2 fs. The particle mesh-Ewald method was used to calculate long-range electrostatic forces, with a mesh density of $1/\text{Å}^3$. Van der Waals forces were calculated with a cut-off of 12 Å and a switching function starting at 10 Å. Periodic boundary conditions were imposed in all MD runs. Temperature was held at 300 K via Langevin dynamics with a damping coefficient of 1 ps^{-1}. Pressure was held at 1 atm using the Nosé-Hoover Langevin piston method with damping time of 50 fs and period of 200 fs.

The system was minimized over 3.4×10^5 steps. All atoms, except for those of the lipid tails, were fixed during the subsequent 4.6 ns equilibration to allow the lipid tails to melt. Lipid head group constraints were then released and equilibration continued for another 10 ns, to allow the membrane to form a watertight seal around the ecMscS pore. The system was then put through 10 cycles of alternating runs, the first being un-constrained equilibration with voltage 0.6 V across the membrane (measured from the cytoplasmic side), and the other runs being equilibration, with no applied voltage, and with ecMscS backbone atoms being harmonically constrained with spring constant 2 kcal/mol/Å^2 to their last positions in the previous run. The runs with applied voltage induced widening of the ecMscS pore [176] while the runs with constraints on the ecMscS backbone allowed the membrane to relax back to a stable state so that the strong voltage did not break up the membrane before the ecMscS pore widened. The resulting structure was used for the production run, during which protein backbone atoms were again put under harmonic restraint with spring constant 2 kcal/mol/Å^2, in order to maintain the widened pore width in the absence of an applied voltage.

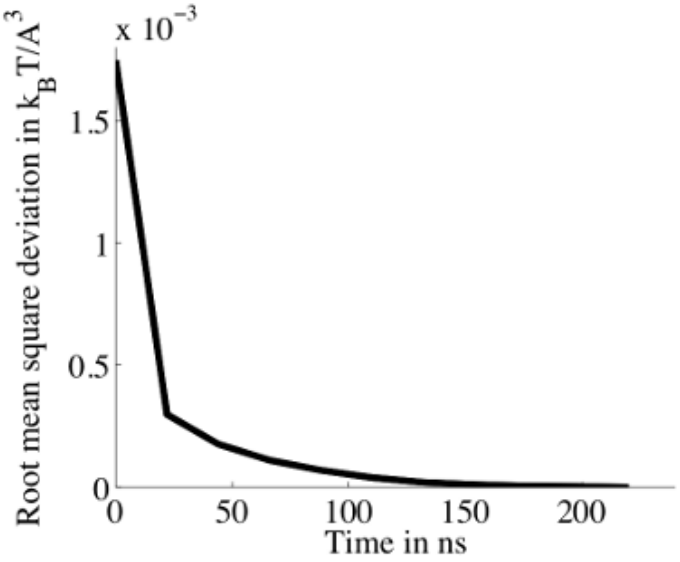

Figure 5.20: Root mean square deviation of the sampled PMF from the final map in a 84 Å × 84 Å × 120 Å box enclosing the ecMscS channel.

MD runs for sampling the PMF and other properties were carried out for 240 ns, ensuring adequate sampling as shown in the plot of the convergence metric used in Fig. 5.20. The convergence metric is defined as the root mean square deviation of the potential map at 20-ns intervals from the final map during the run. Since the interior of ecMscS is the least accessible, and hence the least sampled, region of the system, we confine the convergence metric calculation to a box which just encloses the protein.

5.6.3 Simulation Procedure and Results

Particles were initialized in randomly chosen cells on the cytoplasmic side. The random selection was performed with weights proportional to the volume of each cell, so that the particles were initially uniformly distributed. The particles were then allowed to diffuse. When a particle crosses via ecMscS from the cytoplasmic side to the periplasmic side of the membrane, it is assumed to diffuse away from the membrane and, accordingly, is removed from the system. For the purpose of assessing the effect of transmembrane potential biases, the prior MD simulations were performed with positive, negative, and zero biases. The strengths of the voltage biases followed the study of Sotomayor *et al* [250], which is used as a reference for comparison of results. For each voltage bias, an MD simulation was carried out to obtain the requisite PMF map. The PMF maps were each applied to the present

method as described below.

The simulation procedure was repeated for six runs - two ion species, each with bias voltages 0 mV, +100 mV, −100 mV (measured from the cytoplasmic side). For each simulation, the initial particle count was 5000, and the total simulation time was 4 μs. The number of conduction events for each run is listed in Table 5.2, which shows the results of the present study, scaled to match the initial particle concentration and simulation time of the reference study, together with the results for two putative open conformations of ecMscS in the reference study [250].

In agreement with the reference study, there were few conduction events for K^+ with bias voltages 0 mV and −100 mV. For Glu^-, the numbers of events were much smaller than in the reference study where actually Cl^- ions were used.

Table 5.1: Number of conduction events for each ion for different biasing voltages.

Ion	0 mV	+100 mV	-100 mV
K^+	222	1992	34
Glu^-	217	54	20

Table 5.2: Comparison between present and reference study. Values for the present study have been scaled to account for the different initial particle concentration and simulation time in the reference study. The present study employed Glu^- ions, whereas the reference study employed Cl^- ions. The reference study employed two distinct putative open conformations of the channel. For each respective ion and bias, the results for both conformations are presented, separated by a comma.

	Present			Reference		
Ion	0 mV	+100 mV	-100 mV	0 mV	+100 mV	-100 mV
K^+	2	20	0	2 , 5	13 , 23	3 , 0
Glu^-	2	0	0	22 , 40	6 , 17	53 , 72

5.7 Discussion of Results

The results for K^+ in the present study agree with those of the reference study, while those of Glu^- differ significantly. The significantly lower event count for the anion is attributed to the fact that the present study employed bulky Glu^- ions as compared to the Cl^- ions employed in the reference study. Glu^- ions are actually the prevalent negative ions in *E.*

coli, which is why they have been employed here. During the MD calculations, not only would the bulk diffusion coefficient of the anion be lower in the present study than in the reference one as reported [251] (the diffusion coefficient of Cl^- is similar to that of K^+), but also the resulting steric exclusion in narrow regions around ecMscS results in higher potential barriers in the Glu^- PMF maps.

Other factors contribute to the discrepancies between the results of the present study and the reference study. An examination of the PMF maps shown in Fig. 5.19 reveals potential wells in the vicinity of the ecMscS structure. Ions congregating in these wells present an obstacle to other ions that would otherwise also enter the wells. However, in the absence of inter-ionic interactions, the lack of steric exclusion and local electrostatic interactions in the present version of the kinetic model allows the ions to all linger in the wells, substantially increasing the time taken to reach their designated targets. Such an effect would have been avoided in the reference study because inter-ionic interactions were included in the respective simulations.

The absence of steric effects in the kinetic simulation illustrates the issues arising from the absence of inter-ionic interactions in simulations, especially in regions where ions come into close proximity of one another, such as in the channel interior. Adding inter-ionic interactions in a manner that is both physically sensible and computationally feasible is difficult because such interactions are modulated by environmental factors as well as by the presence of more than two particles within interaction range.

In light of the challenging nature of an account of ion-ion interactions, we propose as a first step a naive solution. One starts by identifying local regions in the system with roughly similar environments. One such region might be the interior of the ecMscS structure, namely the interior of the pore and the CD, and a second might be everywhere outside it. For each region, one determines then the pair correlation function $g(r)$ for the various ion pairs, $K^+ - K^+$, $Glu^- - Glu^-$ and $K^+ - Glu^-$, from the MD trajectories used for the PMF extraction. $g(r)$ can then be used to modulate the transition matrix probabilities of particles that move within a pre-set interaction range of each other.

Another issue of concern is the handling of diffusion coefficients. The assumption in the present study is that the diffusion coefficient for each ion species is constant throughout the system. This assumption was made for the sake of simplicity. However, one expects that the diffusion coefficient of glutamate in the crowded interior of the CD is very different from that in bulk solvent outside of the CD. The proposed remedy is to average over diffusion lengths of ions in local regions of the systems throughout the MD trajectories and from these

lengths obtain the local diffusion coefficient in each grid cell.

It would also behoove us to ensure that the constant-value approximation of the diffusion coefficient, and other quantities for that matter, within each grid cell is valid. For that purpose, one could either interpolate the PMF and diffusion coefficient maps within the cell, or use a sufficiently fine grid to describe regions in which the maps vary sharply. The former would be difficult to implement within the present framework due to the complexity of the additional computation required. The latter can conceivably be a future addition to the algorithm that calculates, for each region, the grid density that resolves the local gradient of maps such that the error between the approximation and the actual quantities fall below a pre-set threshold. Since the PMF map describes the geometry of the system, such a scheme would also be a natural means of quantifying the suitability of grid density to the geometric intricacy of the system.

Since the utility of the present method relies greatly on its computational efficiency, it would be useful to consider alternative ways of solving the rate equation (5.18). In particular, one could consider employing a Chebyshev expansion [252] to approximate the solution

$$\mathbf{P}(t) = e^{\mathbf{R}t} \quad . \tag{5.69}$$

Incorporating such a scheme into the solution of either the rate equation directly or of the local rate matrices one can compare the resulting efficiency with the present method.

5.8 Conclusion

The kinetic model of diffusion described in the present study is a feasible means of simulating diffusion in cellular systems over a wide range of length and time scales. Results from validation tests show robust agreement with analytic descriptions over time step sizes and diffusion coefficients typical of biomolecular systems. Furthermore, the method is adaptable to a wide range of scientific needs and computational capabilities, through the adjustment of simulation parameters. The method can be made more efficient through parallelization of the algorithm and is viable for both deterministic and stochastic calculations.

Algorithmic benefits of the method include accuracy that increases with time step and the restriction of calculations to a local region around each state. These benefits were brought to bear in the ecMscS example, where a simulation on the length scale of hundreds of Ångstroms and a time scale of 1 μs was run serially and completed in two days. The ecMscS example

produced agreement with the reference study in the case of K^+. The examples presented also illustrate the weaknesses of our method, which serve as pointers for future development, namely the inclusion of dielectric effects, the use of position-dependent diffusion coefficients and the inclusion of inter-ion interactions.

REFERENCES

[1] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. i. general method," *J. Chem. Phys.*, vol. 31, pp. 459–466, 1959.

[2] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics II. behavior of a small number of elastic spheres," *J. Chem. Phys.*, vol. 33, pp. 1439–1451, 1960.

[3] J. D. Bernal, "The Bakerian lecture, 1962: The structure of liquids," *Proc. R. Soc.*, vol. 280, no. 1382, pp. 299–322, 1964.

[4] N. M. AB, "The Nobel Prize in Chemistry 2013," 2013. [Online]. Available: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/

[5] J. N. Israelachvili, *Intermolecular and Surface Forces.* London: Academic Press, 1992.

[6] N. Foloppe and A. D. MacKerrell Jr., "All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data," *J. Comp. Chem.*, vol. 21, pp. 86–104, 2000.

[7] C. N. Schutz and A. Warshel, "What are the dielectric "constants" of proteins and how to validate electrostatic models?" *Proteins*, vol. 44, p. 400, 2001.

[8] J. W. Ponder and D. A. Case, "Interatomic potentials and their relative parameters for protein simulations," *Adv. Prot. Chem.*, vol. 66, pp. 27–95, 2003.

[9] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *J. Comp. Chem.*, vol. 25, pp. 1157–1174, 2004.

[10] A. Warshel, P. K. Sharma, M. Kato, and W. W. Parson, "Modeling electrostatic effects in proteins," *Biochim. Biophys. Acta*, vol. 1764, pp. 1647–1676, 2006.

[11] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the gromos force-field versions 54a7 and 54b7." *Eur. Biophys. J.*, vol. 40, pp. 843–856, 2011.

[12] F. Tama, O. Miyashita, and C. L. B. III, "Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM," *J. Struct. Biol.*, vol. 147, no. 3, pp. 315–326, 2004.

[13] K. Suhre, J. Navaza, and Y. H. Sanejouand, "NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps," *Acta Cryst. D*, vol. 62, pp. 1098–1100, 2006.

[14] G. F. Schröder, A. T. Brunger, and M. Levitt, "Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution," *Structure*, vol. 15, pp. 1630–1641, 2007.

[15] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali, "Protein structure fitting and refinement guided by cryo-EM density," *Structure*, vol. 16, pp. 295–307, 2008.

[16] C. C. Jolley, S. A. Wells, P. Fromme, and M. F. Thorpe, "Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations," *Biophys. J.*, vol. 94, pp. 1613–1621, 2008.

[17] P. Afonine, J. Headd, T. Terwilliger, and P. Adams, "New tool: *phenix.real_space_refine*," *Computational Crystallography Newsletter*, vol. 4, no. 2, pp. 43–44, 2013.

[18] J. R. Lopéz-Blanco and P. Chacón, "iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates," *J. Struct. Biol.*, vol. 184, no. 2, pp. 261–270, 2013.

[19] X. Wu, S. Subramaniam, D. A. Case, K. W. Wu, and B. R. Brooks, "Targeted conformational search with map-restrained self-guided langevin dynamics: Application to flexible fitting into electron microscopic density maps," *J. Struct. Biol.*, vol. 183, no. 3, pp. 429–440, 2013.

[20] F. DiMaio, Y. Song, X. Li, M. J. Brunner, C. Xu, V. Conticello, E. Egelman, T. C. Marlovits, Y. Cheng, and D. Baker, "Atomic-accuracy models from 4.5 Å cryo-electron microscopy data with density-guided iterative local refinement," *Nat. Methods*, 2015.

[21] R. McGreevy, I. Teo, A. Singharoy, and K. Schulten, "Advances in the molecular dynamics flexible fitting method for cryo-EM modeling," *Methods*, vol. 100, pp. 50–60, 2016.

[22] A. Merk, A. Bartesaghi, S. Banerjee, V. Falconieri, R. Prashant, M. I. Davis, R. Pragani, M. B. Boxer, L. A. Earl, J. L. S. Milne, and S. Subramaniam, "Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery," *Cell*, vol. 165, no. 7, pp. 1698–1707, 2016.

[23] F. Cérou and A. Guyader, "Adaptive multilevel splitting for rare event analysis," *Stoch. Anal. Appl.*, vol. 25, no. 2, pp. 417–443, 2007.

[24] F. Cérou, A. Guyader, T. Leliévre, and D. Pommier, "A multiple replica approach to simulate reactive trajectories," *J. Chem. Phys.*, vol. 134, no. 5, p. 054108, 2011.

[25] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *J. Comput. Phys.*, vol. 23, no. 2, pp. 187–199, 1977.

[26] B. Ensing, M. D. Vivio, Z. Liu, P. Moore, and M. Klein, "Metadynamics as a tool for exploring free energy landscapes of chemical reactions," *Acc. Chem. Res.*, vol. 39, no. 2, pp. 73–81, 2006.

[27] J. C. Gumbart, B. Roux, and C. Chipot, "Efficient determination of protein-protein standard binding free energies from first principles," *J. Chem. Theory Comput.*, vol. 9, no. 8, pp. 3789–3798, 2013.

[28] C. Chipot and A. Pohorille, Eds., *Free Energy Calculations. Theory and Applications in Chemistry and Biology.* Berlin, Germany: Springer Verlag, 2007.

[29] H. Kahn and T. E. Harris, "Estimation of particle transmission by random sampling," *National Bureau of Standards Appl. Math. Series*, vol. 12, pp. 27–30, 1951.

[30] T. S. van Erp and P. G. Bolhuis, "Elaborating transition interface sampling methods," *J. Comp. Phys.*, vol. 205, no. 1, pp. 157–181, 2005.

[31] R. J. Allen, C. Valeriani, and P. R. ten Wolde, "Forward flux sampling for rare event simulations," *J. Phys.-Condens. Mat.*, vol. 21, no. 46, p. 463102, 2009.

[32] D. L. Ermak, "A computer simulation of charged particles in solution. I. Technique and equilibrium properties," *J. Chem. Phys.*, vol. 62, no. 10, pp. 4162–4196, 1975.

[33] S. R. McGuffee and A. H. Elcock, "Atomically detailed simulations of concentrated protein solutions: the effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems," *J. Am. Chem. Soc.*, vol. 128, no. 37, pp. 12 098–12 110, 2006.

[34] T. Frembgen-Kesner and A. H. Elcock, "Striking effects of hydrodynamic interactions on the simulated diffusion and folding of proteins," *J. Chem. Theor. Comp.*, vol. 5, no. 2, pp. 242–256, 2009.

[35] T. Ando and J. Skolnick, "Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 43, pp. 18 457–18 462, 2010.

[36] T. van der Straaten, G. Kathawala, and U. Ravaioli, "Biomoca: A transport Monte Carlo model for ion channels," *J. Comp. Electron.*, vol. 2, no. 2-4, pp. 231–237, 2003.

[37] A. V. Popov and N. Agmon, "Three-dimensional simulations of reversible bimolecular reactions: The simple target problem," *J. Chem. Phys.*, vol. 115, no. 19, p. 8921, 2001.

[38] J. S. van Zon and P. R. ten Wolde, "Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space," *J. Chem. Phys.*, vol. 123, no. 23, p. 234910, 2005.

[39] J. S. van Zon and P. R. ten Wolde, "Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics," *Phys. Rev. Lett.*, vol. 94, p. 128103, 2005.

[40] R. Carr, J. Comer, M. D. Ginsberg, and A. Aksimentiev, "Atoms-to-microns model for small solute transport through sticky nanochannels," *Lab Chip*, 2011, dOI:10.1039/C1LC20697D.

[41] A. Rahman, "Correlations in the motion of atoms in liquid argon," *Phys. Rev. A*, vol. 136, pp. 405–411, 1964.

[42] M. Karplus, "Molecular dynamics: Applications to proteins," in *Modelling of Molecular Structures and Properties*, ser. Studies in Physical and Theoretical Chemistry, J.-L. Rivail, Ed., vol. 71. Amsterdam: Elsevier Science Publishers, 1990, proceedings of an International Meeting. pp. 427–461.

[43] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, "A coarse grain model for phospholipid simulations," *J. Phys. Chem. B*, vol. 105, pp. 4464–4470, 2001.

[44] F. Müller-Plathe, "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," *ChemPhysChem*, vol. 3, pp. 754–769, 2002.

[45] M. Müller, K. Katsov, and M. Schick, "Coarse-grained models and collective phenomena in membranes: Computer simulation of membrane fusion," *J. Polym. Sci. B*, vol. 41, pp. 1441–1450, 2003.

[46] S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," *J. Phys. Chem. B*, vol. 108, pp. 750–760, 2004.

[47] V. Tozzini and J. A. McCammon, "A coarse grained model for the dynamics of flap opening in the hiv-1 protease," *Chem. Phys. Lett.*, vol. 413, no. 1-3, pp. 123–128, 2005.

[48] A. Y. Shih, A. Arkhipov, P. L. Freddolino, and K. Schulten, "Coarse grained protein-lipid model with application to lipoprotein particles," *J. Phys. Chem. B*, vol. 110, pp. 3674–3684, 2006.

[49] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI force field: coarse grained model for biomolecular simulations," *J. Phys. Chem. B*, vol. 111, pp. 7812–7824, 2007.

[50] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *J. Am. Chem. Soc.*, vol. 112, pp. 6127–6129, 1990.

[51] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 18, pp. 10 037–10 041, 2001.

[52] L. Onufriev, D. A. Case, and D. Bashford, "Effective Born Radii in the Generalized Born Approximation: The Importance of Being Perfect," *J. Comp. Chem.*, vol. 23, pp. 1297–1304, 2002.

[53] W. Im, M. Feig, and C. L. Brooks III, "An implicit membrane generalized Born theory for the study of structure, stability, and interactions of membrane proteins," *Biophys. J.*, vol. 85, pp. 2900–2918, 2003.

[54] B. Lu, D. Zhang, and J. A. McCammon, "Computation of electrostatic forces between solvated molecules determined by the poisson-boltzmann equation using a boundary element method," *J. Chem. Phys.*, vol. 122, no. 21, p. 214102, 2005.

[55] D. E. Tanner, K.-Y. Chan, J. Phillips, and K. Schulten, "Parallel generalized Born implicit solvent calculations with NAMD," *J. Chem. Theor. Comp.*, vol. 7, pp. 3635–3642, 2011.

[56] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, I. W. E. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, pp. 3586–3616, 1998.

[57] J. E. Lennard-Jones, "On the determination of molecular fields," *Proc. R. Soc. Lond. A. (Math. Phys. Sci.)*, vol. 106, no. 738, pp. 436–477, 1924.

[58] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles," *J. Chem. Theor. Comp.*, vol. 8, no. 9, pp. 3257–3273, 2012. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct300400x

[59] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, 1995.

[60] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," *J. Comp. Chem.*, vol. 25, pp. 1656–1676, 2004.

[61] V. M. Unger, "Electron cryomicroscopy methods," *Curr. Opin. Struct. Biol.*, vol. 11, no. 5, pp. 548–554, 2002.

[62] R. Neutze, G. Brändén, and G. F. Schertler, "Membrane protein structural biology using x-ray free electron lasers," *Curr. Opin. Struct. Biol.*, vol. 33, pp. 115–125, 2015.

[63] J. A. Velazquez-Muriel, M. Valle, A. Santamaría-Pang, I. A. Kakadiaris, and J. M. Carazo, "Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies," *Structure*, vol. 14, pp. 1115–1126, 2006.

[64] M. Orzechowski and F. Tama, "Flexible fitting of high-resolution X-ray structures into cryo electron microscopy maps using biased molecular dynamics simulations," *Biophys. J.*, vol. 95, no. 12, pp. 5692–5705, 2008.

[65] J. A. Kovacs, M. Yeager, and R. Abagyan, "Damped-dynamics flexible fitting," *Biophys. J.*, vol. 95, pp. 3192–3207, 2008.

[66] L. G. Trabuco, E. Villa, K. Mitra, J. Frank, and K. Schulten, "Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics," *Structure*, vol. 16, pp. 673–683, 2008.

[67] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten, "Molecular Dynamics Flexible Fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography," *Methods*, vol. 49, pp. 174–180, 2009.

[68] B. C. Goh, J. R. Perilla, M. R. England, K. J. Heyrana, R. C. Craven, and K. Schulten, "Atomic modeling of an immature retroviral lattice using molecular dynamics and mutagenesis," *Structure*, vol. 23, pp. 1414–1425, 2015.

[69] E. Villa, J. Sengupta, L. G. Trabuco, J. LeBarron, W. T. Baxter, T. R. Shaikh, R. A. Grassucci, P. Nissen, M. Ehrenberg, K. Schulten, and J. Frank, "Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 1063–1068, 2009.

[70] L. G. Trabuco, E. Schreiner, J. Gumbart, J. Hsin, E. Villa, and K. Schulten, "Applications of the molecular dynamics flexible fitting method," *J. Struct. Biol.*, vol. 173, pp. 420–427, 2011.

[71] J. Frauenfeld, J. Gumbart, E. O. van der Sluis, S. Funes, M. Gartmann, B. Beatrix, T. Mielke, O. Berninghausen, T. Becker, K. Schulten, and R. Beckmann, "Cryo-EM structure of the ribosome-SecYE complex in the membrane environment," *Nat. Struct. Mol. Biol.*, vol. 18, pp. 614–621, 2011.

[72] S. Wickles, A. Singharoy, J. Andreani, S. Seemayer, L. Bischoff, O. Berninghausen, J. Soeding, K. Schulten, E. van der Sluis, and R. Beckmann, "A structural model of the active ribosome-bound membrane protein insertase YidC," *eLIFE*, vol. 3:e03035, 2014, (17 pages).

[73] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang, "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics," *Nature*, vol. 497, pp. 643–646, 2013.

[74] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng, "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM," *Nat. Methods*, vol. 10, pp. 584–590, 2013.

[75] A.-C. Milazzo, A. Cheng, A. Moeller, D. Lyumkis, E. Jacovetty, J. Polukas, M. H. Ellisman, N.-H. Xuong, B. Carragher, and C. S. Potter, "Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy." *J. Struct. Biol.*, vol. 176, no. 3, pp. 404–8, Dec. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3210420\&tool=pmcentrez\&rendertype=abstract

[76] M. Liao, E. Cao, D. Julius, and Y. Cheng, "Structure of the TRPV1 ion channel determined by electron cryo-microscopy," *Nature*, vol. 504, pp. 107–112, 2013.

[77] A. Bartesaghi, D. Matthies, S. Banerjee, A. Merk, and S. Subramaniam, "Structure of β-galactosidase at 3.2 Å resolution obtained by cryo-electron microscopy," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 32, pp. 11 709–11 714, 2014.

[78] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, J. L. Milne, and S. Subramaniam, "2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor," *Science*, vol. 348, no. 6239, pp. 1147–1151, 2015.

[79] M. Zhao, S. Wu, Q. Zhou, S. Vivona, D. J. Cipriano, Y. Cheng, and A. T. Brunger, "Mechanistic insights into the recycling machine of the SNARE complex," *Nature*, vol. 518, pp. 61–67, 2015.

[80] N. Fischer, P. Neumann, A. L. Konevega, L. V. Bock, R. Ficner, M. V. Rodnina, and H. Stark, "Structure of the *E. coli* ribosome-EF-Tu complex at <3 Å resolution by c_s-corrected cryo-EM," *Nature*, vol. 520, pp. 567–570, 2015.

[81] A. Brown, S. Shao, J. Murray, R. S. Hegde, and V. Ramakrishnan, "Structural basis for stop codon recognition in eukaryotes," *Nature*, 2015, in press.

[82] R. Samudrala, Y. Xia, E. S. Huang, and M. Levitt, "Ab initio prediction of protein structure using a combined hierarchical approach," *Proteins: Structure, Function, and Genetics*, vol. S3, pp. 194–198, 1999.

[83] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field," *Proteins*, vol. 80, pp. 1715–1735, 2012.

[84] A. Singharoy, B. Venkatakrishnan, Y. Liu, C. G. Mayne, C.-H. Chen, A. Zlotnick, K. Schulten, and A. H. Flood, "Macromolecular crystallography for synthetic abiological molecules: Combining xMDFF and PHENIX for structure determination of cyanostar macrocycles," *J. Am. Chem. Soc.*, vol. 137, pp. 8810–8818, 2015.

[85] B. A. Barad, N. Echols, R. Y.-R. Wang, Y. Cheng, F. DiMaio, P. D. Adams, and J. S. Fraser, "Emringer: side chain-directed model and map validation for 3d cryo-electron microscopy," *Nat. Methods*, vol. 12, no. 10, pp. 943–946, 2015.

[86] V. B. Chen, W. B. A. III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Cryst. D*, vol. 66, pp. 12–21, 2010.

[87] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *J. Mol. Graphics*, vol. 14, pp. 33–38, 1996.

[88] W. Wriggers, "Using situs for the integration of multi-resolution structures," *Biophysical Reviews*, vol. 2, pp. 21–27, 2010.

[89] J. Hsin, J. Gumbart, L. G. Trabuco, E. Villa, P. Qian, C. N. Hunter, and K. Schulten, "Protein-induced membrane curvature investigated through molecular dynamics flexible fitting," *Biophys. J.*, vol. 97, pp. 321–329, 2009.

[90] M. K. Sener, J. Hsin, L. G. Trabuco, E. Villa, P. Qian, C. N. Hunter, and K. Schulten, "Structural model and excitonic properties of the dimeric RC-LH1-PufX complex from *Rhodobacter sphaeroides*," *Chem. Phys.*, vol. 357, pp. 188–197, 2009.

[91] J. Gumbart, C. Chipot, and K. Schulten, "Free energy of nascent-chain folding in the translocon," *J. Am. Chem. Soc.*, vol. 133, no. 19, pp. 7602–7607, 2011.

[92] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chem. Phys. Lett.*, vol. 314, pp. 141–151, 1999.

[93] W. Jiang, J. Phillips, L. Huang, M. Fajer, Y. Meng, J. Gumbart, Y. Luo, K. Schulten, and B. Roux, "Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD," *Comp. Phys. Commun.*, vol. 185, pp. 908–916, 2014.

[94] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comp. Chem.*, vol. 29, pp. 1859–1865, 2008.

[95] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell Jr., and R. W. Pastor, "Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types," *J. Phys. Chem. B*, vol. 114, no. 23, pp. 7830–7843, 2010.

[96] T. R. Shaikh, H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith, and J. Frank, "SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs ," *Nat. Protoc.*, vol. 3, no. 12, pp. 1941–1974, 2008.

[97] J. E. Stone, J. Gullingsrud, P. Grayson, and K. Schulten, "A system for interactive molecular dynamics simulation," in *2001 ACM Symposium on Interactive 3D Graphics*, J. F. Hughes and C. H. Séquin, Eds. New York: ACM SIGGRAPH, 2001, pp. 191–194.

[98] P. Grayson, E. Tajkhorshid, and K. Schulten, "Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics," *Biophys. J.*, vol. 85, pp. 36–48, 2003.

[99] R. McGreevy, A. Singharoy, Q. Li, J. Zhang, D. Xu, E. Perozo, and K. Schulten, "xMDFF: Molecular dynamics flexible fitting of low-resolution X-Ray structures," *Acta Cryst. D*, vol. 70, pp. 2344–2355, 2014.

[100] C. Darnault, A. Volbeda, E. J. Kim, P. Legrand, X. Vernède, P. A. Lindahl, and J. C. Fontecilla-Camps, "Ni-Zn-[Fe_4-S_4] and Ni-Ni-[Fe_4-S_4] clusters in closed and open α subunits of acetyl-CoA synthase/carbon monoxide dehydrogenase," *Nat. Struct. Biol.*, vol. 10, pp. 271–279, 2003.

[101] P. Adams, P. Afonine, G. Bunkóczi, V. Chen, I. Davis, N. Echols, J. Headd, L. Hung, G. Kapral, R. Grosse-Kunstleve, A. McCoy, N. Moriarty, R. Oeffner, R. Read, D. Richardson, J. Richardson, T. Terwilliger, and P. Zwart, "PHENIX: a comprehensive Python-based system for macromolecular structure solution." *Acta Cryst. D*, vol. 66, pp. 213–221, Feb. 2010.

[102] J. E. Stone, R. McGreevy, B. Isralewitz, and K. Schulten, "GPU-accelerated analysis and visualization of large structures solved by molecular dynamics flexible fitting," *Faraday Discuss.*, vol. 169, pp. 265–283, 2014.

[103] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler et al., "Rosetta3: an object-oriented software suite for the simulation and design of macromolecules," *Meth. Enzym.*, vol. 487, p. 545, 2011.

[104] A. Brunger, P. D. Adams, P. Fromme, R. Fromme, M. Levitt, and G. F. Schröder, "Improving the accuracy of macromolecular structure refinement at 7 Å resolution," *Structure*, vol. 20, pp. 957–966, June 2012.

[105] G. F. Schröder, M. Levitt, and A. Brunger, "Super-resolution biomolecular crystallography with low-resolution data," *Nature*, vol. 464, pp. 1218–1222, Apr. 2010.

[106] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, "Quantifying the local resolution of cryo-EM density maps," *Nat. Methods*, vol. 11, no. 1, pp. 63–65, 2014.

[107] A. E. Leschziner and E. Nogales, "Visualizing flexibility at molecular resolution: Analysis of heterogeneity in single-particle electron microscopy reconstructions," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, pp. 43–62, 2007.

[108] A. Singharoy, I. Teo, R. McGreevy, J. E. Stone, J. Zhao, and K. Schulten, "Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps ," *eLIFE*, vol. 5, p. e16105, 2016.

[109] H. N. C. *et al*, "Femtosecond X-ray protein nanocrystallography," *Nature*, vol. 470, pp. 73–77, 2013.

[110] L. D. Landau, *Statistical Physics*. Oxford, UK: The Clarendon Press, 1938.

[111] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," *J. Chem. Phys.*, vol. 22, no. 8, pp. 1420–1426, 1954.

[112] Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations," *J. Phys. Chem. B*, vol. 113, no. 8, pp. 2234–2246, 2009.

[113] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 20, pp. 12 562–12 566, 2002.

[114] E. Darve and A. Pohorille, "Calculating free energies using average force," *J. Chem. Phys.*, vol. 115, no. 20, pp. 9169–9183, 2001.

[115] B. Roux, "The calculation of the potential of mean force using computer simulations," *Comput. Phys. Commun.*, vol. 91, no. 1-3, pp. 275–282, 1995.

[116] C. Bartels, M. Schaeffer, and M. Karplus, "Determination of equilibrium properties of biomolecular systems using multidimensional adaptive umbrella sampling," *J. Chem. Phys.*, vol. 111, no. 17, pp. 8048–8067, 1999.

[117] C. Neale, T. Rodinger, and R. Pomès, "Equilibrium exchange enhances the convergence rate of umbrella sampling," *Chem. Phys. Lett.*, vol. 460, no. 1-3, pp. 375–381, 2008.

[118] W. Jiang and B. Roux, "Free energy perturbation hamiltonian replica-exchange molecular dynamics (fep/h-remd) for absolute ligand binding free energy calculations," *J. Chem. Theory Comput.*, vol. 6, no. 9, pp. 2559–2565, 2010.

[119] M. Moradi and E. Tajkhorshid, "Driven metadynamics: Reconstructing equilibrium free energies from driven adaptive-bias simulations," *J. Phys. Chem. Lett.*, vol. 4, no. 11, pp. 1882–1887, 2013.

[120] D. S. Dashti and A. E. Roitberg, "Optimization of umbrella sampling replica exchange molecular dynamics by replica positioning," *J. Chem. Theory Comput.*, vol. 9, no. 11, pp. 4692–4699, 2013.

[121] D. I. Kopelevich, "One-dimensional potential of mean force underestimates activation barrier for transport across flexible lipid membranes," *J. Chem. Phys.*, vol. 139, no. 13, p. 134906, 2013.

[122] T. S. van Erp, D. Moroni, and P. G. Bolhuis, "A novel path sampling method for the calculation of rate constants," *J. Chem. Phys.*, vol. 118, no. 17, pp. 7762–7774, 2003.

[123] R. J. Allen, P. B. Warren, and P. R. ten Wolde, "Sampling rare switching events in biochemical networks," *Phys. Rev. Lett.*, vol. 94, p. 018104, 2005.

[124] H. Lu and P. J. Tonge, "Drug-target residence time: critical information for lead optimization," *Curr. Opin. Chem. Biol.*, vol. 14, no. 4, pp. 467–474, 2010.

[125] R. A. Copeland, "The drug-target residence time model: a 10-year retrospective," *Nat. Rev. Drug Discov.*, vol. 15, pp. 87–95, 2015.

[126] D. Huang and A. Caflisch, "The free energy landscape of small molecule unbinding," *PLoS Comput. Biol.*, vol. 7, no. 2, p. e1002002, 2011.

[127] I. Buch, T. Giorgino, and G. D. Fabriitis, "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 25, pp. 10 184–10 189, 2011.

[128] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello, "Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps," *P. Natl. Acad. Sci. USA*, vol. 112, no. 5, pp. E386–E391, 2014.

[129] N. Plattner and F. Noé, "Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models," *Nat. Commun.*, vol. 6, p. 7653, 2015.

[130] F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules," *Curr. Opin. Struct. Biol.*, vol. 18, no. 2, pp. 154–162, 2008.

[131] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: generation and validation," *J. Chem. Phys.*, vol. 134, no. 17, p. 174105, 2011.

[132] C.-E. Bréhier, M. Gazeau, L. Goudenège, T. Lelièvre, and M. Rousset, "Unbiasedness of some generalized adaptive multilevel splitting algorithms," *arXiv:1505.02674 [math.PR]. arXiv.org ePrint archive*, 2015. [Online]. Available: http://arxiv.org/abs/1505.02674

[133] A. Guyader, N. Hengartner, and E. Matzner-Løber, "Simulation and estimation of extreme quantiles and extreme probabilities," *Appl. Math. Optim.*, vol. 64, no. 2, pp. 171–196, 2011.

[134] F. Cérou, P. D. Moral, T. Furon, and A. Guyader, "Sequential monte carlo for rare event estimation," *Stat. Comput.*, vol. 22, no. 3, pp. 795–808, 2012.

[135] D. Aristoff, T. Lelièvre, C. G. Mayne, and I. Teo, "Adaptive multilevel splitting in molecular dynamics simulations," *ESAIM Proc. Surv.*, vol. 48, pp. 215–225, 2015.

[136] C. E. Bréhier, T. Lelièvre, and M. Rousset, "Analysis of adaptive multilevel splitting algorithms in an idealized case," *ESAIM Proc. Surv.*, vol. 19, pp. 361–394, 2015.

[137] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005.

[138] D. Beglov and B. Roux, "Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations," *J. Chem. Phys.*, vol. 100, no. 12, pp. 9050–9063, 1994.

[139] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[140] A. Szabo, K. Schulten, and Z. Schulten, "First passage time approach to diffusion controlled reactions," *J. Chem. Phys.*, vol. 72, no. 8, pp. 4350–4356, 1980.

[141] T. B. Woolf and B. Roux, "Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: A molecular dynamics study of solvation effects," *J. Am. Chem. Soc.*, vol. 116, no. 13, pp. 5916–5926, 1994.

[142] G. Hummer, "Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations," *New J. Phys.*, vol. 7, no. 1, p. 34, 2005.

[143] J. W. Essex, D. L. Severance, J. Tirado-Rives, and W. L. Jorgensen, "Monte carlo simulations for proteins: Binding affinities for trypsin-benzamidine complexes via free energy perturbations," *J. Phys. Chem. B*, vol. 101, no. 46, pp. 9663–9669, 1997.

[144] H. Resat, T. J. Marrone, and J. A. McCammon, "Enzyme-inhibitor association thermodynamics: Explicit and continuum solvent studies," *Biophys. J.*, vol. 72, no. 2 Pt 1, pp. 552–532, 1997.

[145] G. Guillain and D. Thusius, "Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin," *J. Am. Chem. Soc.*, vol. 92, no. 18, pp. 5534–5536, 1970.

[146] H. Lu, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten, "Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation," *Biophys. J.*, vol. 75, no. 2, pp. 662–671, 1998.

[147] A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *Q. J. R. Meteorol. Soc.*, vol. 81, no. 348, pp. 158–171, 1955.

[148] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Math. Proc. Cambridge Philos. Soc.*, vol. 24, no. 2, pp. 180–190, 1928.

[149] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, 1983.

[150] J. Huang and A. D. M. Jr., "Charmm36 all-atom additive protein force field: validation based on comparison to nmr data," *J. Comput. Chem.*, vol. 34, no. 25, pp. 2135–2145, 2013.

[151] G. Fiorin, M. L. Klein, and J. Hénin, "Using collective variables to drive molecular dynamics simulations," *Mol. Phys.*, vol. 111, no. 2, pp. 3345–3362, 2013.

[152] W. Jiang, D. J. Hardy, J. C. Phillips, A. D. M. Jr., K. Schulten, and B. Roux, "High-performance scalable molecular dynamics simulations of a polarizable force field based on classical drude oscillators in namd," *J. Phys. Chem. Lett.*, vol. 2, no. 2, pp. 87–92, 2011.

[153] P. E. M. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. M. Jr., "Force field for peptides and proteins based on the classical drude oscillator," *J. Chem. Theory Comput.*, vol. 9, no. 12, pp. 5430–5449, 2013.

[154] H. Lu, K. England, C. am Ende, J. J. Truglio, S. Luckner, B. G. Reddy, N. L. Marlenee, S. E. Knudson, D. L. Knudson, R. A. Bowen, C. Kisker, R. A. Slayden, and P. J. Tonge, "Slow-onset inhibition of the fabi enoyl reductase from francisella tularensis: residence time and in vivo activity," *ACS Chem. Biol.*, vol. 4, no. 3, pp. 221–231, 2009.

[155] D. Guo, T. Mulder-Krieger, A. P. IJzerman, and L. H. Heitman, "Functional efficacy of adenosine a_2a receptor agonists is positively correlated to their receptor residence time," *Br. J. Pharmacol.*, vol. 166, no. 6, pp. 1846–1859, 2012.

[156] R. Peters, A. Brünger, and K. Schulten, "Continuous fluorescence microphotolysis: A sensitive method for study of diffusion processes in single cells," *Proc. Natl. Acad. Sci. USA*, vol. 78, no. 2, pp. 962–966, 1981.

[157] G. Lamm and K. Schulten, "Extended Brownian dynamics approach to diffusion-controlled processes," *J. Chem. Phys.*, vol. 75, no. 1, pp. 365–371, 1981.

[158] R. Gamini, M. Sotomayor, C. Chipot, and K. Schulten, "Cytoplasmic domain filter function in the mechanosensitive channel of small conductance," *Biophys. J.*, vol. 101, no. 1, pp. 80–89, 2011.

[159] H. C. Berg and E. M. Purcell, "Physics of chemoreception," *Biophys. J.*, vol. 20, no. 2, pp. 193–219, 1977.

[160] D. Shoup and A. Szabo, "Role of diffusion in ligand binding to macromolecules and cell-bound receptors," *Biophys. J.*, vol. 40, no. 1, pp. 33–39, 1982.

[161] K. Sharp, R. Fine, K. Schulten, and B. Honig, "Brownian dynamics simulation of diffusion to irregular bodies," *Phys. Chem.*, vol. 91, pp. 3624–3631, 1987.

[162] R. Zwanzig and A. Szabo, "Time dependent rate of diffusion-influenced ligand binding to receptors on cell surfaces," *Biophys. J.*, vol. 60, no. 3, pp. 671–678, 1991.

[163] R. R. Gabdoulline and R. C. Wade, "Biomolecular diffusional association," *Curr. Opin. Struct. Biol.*, vol. 12, no. 2, pp. 204–213, 2002.

[164] A. Atri, J. Amundson, D. Clapham, and J. Sneyd, "A single-pool model for intracellular calcium oscillations and waves in the Xenopus laevis oocyte," *Biophys. J.*, vol. 65, no. 4, pp. 1727–1739, 1993.

[165] J. Elf, A. Dončić, and M. Ehrenberg, "Mesoscopic reaction-diffusion in intracellular signaling," *Proc. SPIE*, vol. 5110, pp. 114–124, 2003.

[166] K.-H. Chiam, C. M. Tan, V. Bhargava, and G. Rajagopal, "Hybrid simulations of stochastic reaction-diffusion processes for modeling intracellular signaling pathways," *Phys. Rev. E*, vol. 74, no. 5, p. 051910, 2006.

[167] A. Camilli and B. L. Bassler, "Bacterial small-molecule signaling pathways," *Science*, vol. 311, no. 5764, pp. 1113–1116, 2006.

[168] J. Sneyd, B. T. Wetton, A. C. Charles, and M. J. Sanderson, "Intercellular calcium waves mediated by diffusion of inositol trisphosphate: a two-dimensional model," *Am. J. Physiol. – Cell Physiol.*, vol. 268, no. 6, pp. C1537–C1545, 1995.

[169] S. Y. Shvartsman, H. S. Wiley, W. M. Deen, and D. A. Lauffenburger, "Spatial range of autocrine signaling: modeling and computational analysis," *Biophys. J.*, vol. 81, no. 4, pp. 1854–1867, 2001.

[170] W. Nadler and K. Schulten, "Generalized moment expansion for brownian relaxation processes," *J. Chem. Phys.*, vol. 82, pp. 151–160, 1985.

[171] S. S. Andrews, N. J. Addy, R. Brent, and A. P. Arkin, "Detailed simulations of cell biology with Smoldyn 2.1," *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000705, 2010.

[172] J. R. Stiles and T. M. Bartol, *Computational Neuroscience: Realistic Modeling for Experimentalists*, E. D. Schutter, Ed. CRC Press, 2001.

[173] A. E. Cowan, I. I. Moraru, J. C. Schaff, B. M. Slepchenko, and L. M. Loew, "Spatial modeling of cell signaling networks," *Methods Cell Biol.*, vol. 110, pp. 195–221, 2012.

[174] M. Compoint, F. Picaud, C. Ramseyer, and C. Giradet, "Targeted molecular dynamics of an open-state KcsA channel," *J. Chem. Phys.*, vol. 122, no. 13, p. 134707, 2005.

[175] T. W. Allen, O. S. Andersen, and B. Roux, "Ion permeation through a narrow channel: Using gramicidin to ascertain all-atom molecular dynamics potential of mean force methodology and biomolecular force fields," *Biophys. J.*, vol. 90, no. 10, pp. 3447–3468, 2006.

[176] M. Sotomayor, V. Vásquez, E. Perozo, and K. Schulten, "Ion conduction through mscs as determined by electrophysiology and simulation," *Biophys. J.*, vol. 92, pp. 886–902, 2007.

[177] R. Elber, "Long-timescale simulation methods," *Curr. Opin. Struct. Biol.*, vol. 15, no. 2, pp. 151–156, 2005.

[178] D. B. Wells, M. Belkin, J. Comer, and A. Aksimentiev, "Assessing graphene nanopores for sequencing DNA," *Nano Lett.*, vol. 12, pp. 4117–4123, 2012.

[179] J. Comer and A. Aksimentiev, "Predicting the DNA sequence dependence of nanopore ion current using atomic-resolution Brownian dynamics," *J. Phys. Chem. C*, vol. 116, pp. 3376–3393, 2012.

[180] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[181] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps: An Introduction*. New York: Addison-Wesley, 1992.

[182] T. M. Martinetz, S. G. Berkovich, and K. Schulten, "'neural-gas' network for vector quantization and its application to time-series prediction," *IEEE Trans. Neur. Netw.*, vol. 4, no. 4, pp. 558–569, 1993.

[183] T. Martinetz and K. Schulten, "Topology representing networks," *Neur. Netw.*, vol. 7, no. 3, pp. 507–522, 1994.

[184] Q. Du, V. Faber, and M. Gunzburger, "Centroidal Voronoi tessellations: Applications and algorithms," *SIAM Rev.*, vol. 41, no. 4, pp. 637–676, 1999.

[185] J. C. Gumbart, I. Teo, B. Roux, and K. Schulten, "Reconciling the roles of kinetic and thermodynamic factors in membrane-protein insertion," *J. Am. Chem. Soc.*, vol. 135, no. 6, pp. 2291–2297, 2013.

[186] J. Madura, J. Briggs, R. Wade, M. Davis, B. Luty, A. Illin, J. Antosiewicz, M. Gilson, B. Bagheri, L. Scott, and J. A. McCammon, "Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program," *Comp. Phys. Commun.*, vol. 91, no. 1-3, pp. 57–95, 1995.

[187] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids.* Amen House, London E.C.4: Oxford University Press, 1959.

[188] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne, "Recognition of transmembrane helices by the endoplasmic reticulum translocon," *Nature*, vol. 433, pp. 377–381, 2005.

[189] T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, L. Nilsson, S. H. White, and G. von Heijne, "Molecular code for transmembrane-helix recognition by the Sec61 translocon," *Nature*, vol. 450, pp. 1026–1030, 2007.

[190] A. Radzicka and R. Wolfenden, "Comparing the polarities of the amino-acids - side-chain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous-solution," *Biochem.*, vol. 27, no. 5, pp. 1664–1670, 1988.

[191] R. Wolfenden, "Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins," *J. Gen. Physiol.*, vol. 129, no. 5, pp. 357–362, 2007.

[192] J. L. MacCallum and D. P. Tieleman, "Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions," *Trends Biochem. Sci.*, vol. 36, no. 12, pp. 653–662, 2011.

[193] D. Shental-Bechor, S. J. Fleishman, and N. Ben-Tal, "Has the code for protein translocation been broken?" *Trends Biochem. Sci.*, vol. 31, no. 4, pp. 192–196, 2006.

[194] A. Rychkova, S. Vicatos, and A. Warshel, "On the energetics of translocon-assisted insertion of charged transmembrane helices into membranes," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 41, pp. 17 598–17 603, 2010.

[195] T. Junne, L. Kocik, and M. Spiess, "The hydrophobic core of the Sec61 translocon defines the hydrophobicity threshold for membrane integration," *Mol. Biol. Cell*, vol. 21, no. 10, pp. 1662–1670, 2010.

[196] B. Zhang and T. F. M. III, "Hydrophobically stabilized open state for the lateral gate of the Sec translocon," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 12, pp. 5399–5404, 2010.

[197] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein, "Exploring multidimensional free energy landscapes using time-dependent biases on collective variables," *J. Chem. Theor. Comp.*, vol. 6, no. 1, pp. 35–47, 2010.

[198] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," *J. Comp. Chem.*, vol. 16, no. 11, pp. 1339–1350, 1995.

[199] E. Park and T. A. Rapoport, "Mechanisms of Sec61/SecY-mediated protein translocation across membranes," *Annu. Rev. Biophys.*, vol. 41, pp. 21–40, 2012.

[200] T. Tsukazaki, H. Mori, S. Fukai, R. Ishitani, T. Mori, N. Dohmae, A. Perederina, Y. Sugita, D. G. Vassylyev, K. Ito, and O. Nureki, "Conformational transition of Sec machinery inferred from bacterial SecYE structures," *Nature*, vol. 455, pp. 988–991, 2008.

[201] J. Zimmer, Y. Nam, and T. A. Rapoport, "Structure of a complex of the atpase seca and the protein-translocation channel," *Nature*, vol. 455, pp. 936–943, 2008.

[202] P. F. Egea and R. M. Stroud, "Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 40, pp. 17 182–17 187, 2010.

[203] S. M. Simon and G. Blobel, "A protein-conducting channel in the endoplasmic reticulum," *Cell*, vol. 65, no. 3, pp. 371–380, 1991.

[204] D. Heritage and W. F. Wonderlin, "Translocon pores in the endoplasmic reticulum are permeable to a neutral, polar molecule," *J. Biol. Chem.*, vol. 276, no. 25, pp. 22 655–22 662, 2001.

[205] W. F. Wonderlin, "Constitutive, translation-independent opening of the protein-conducting channel in the endoplasmic reticulum," *Pflug. Arch. Eur. J. Physiol.*, vol. 457, pp. 917–930, 2009.

[206] J. Gumbart, L. G. Trabuco, E. Schreiner, E. Villa, and K. Schulten, "Regulation of the protein-conducting channel by a bound ribosome," *Structure*, vol. 17, no. 11, pp. 1453–1464, 2009.

[207] S. H. White and G. von Heijne, "How translocons select transmembrane helices," *Annu. Rev. Biophys.*, vol. 37, pp. 23–42, 2008.

[208] T. Hessa, S. H. White, and G. von Heijne, "Membrane insertion of a potassium-channel voltage sensor," *Science*, vol. 307, no. 5714, p. 1427, 2005.

[209] R. Bol, J. G. de Wit, and A. J. M. Driessen, "The active protein-conducting channel of *Escherichia coli* contains an apolar patch," *J. Biol. Chem.*, vol. 282, no. 41, pp. 29 785–29 793, 2007.

[210] B. van den Berg, W. M. C. Jr., I. Collinson, Y. Modis, E. Hartmann, S. C. Harrison, and T. A. Rapoport, "X-ray structure of a protein-conducting channel," *Nature*, vol. 427, pp. 36–44, 2004.

[211] Z. Cheng and R. Gilmore, "Slow translocon gating causes cytosolic exposure of transmembrane and lumenal domains during membrane protein integration," *Nat. Struct. Mol. Biol.*, vol. 13, pp. 930–936, 2006.

[212] R. Young and H. Bremer, "Polypeptide-chain-elongation rate in escherichia coli b/r as a function of growth rate," *Biochem. J.*, vol. 160, no. 2, pp. 185–194, 1976.

[213] T. Hessa, M. Monné, and G. von Heijne, "Stop-transfer efficiency of marginally hydrophobic segments depends on the length of the carboxy-terminal tail," *EMBO Rep.*, vol. 4, no. 2, pp. 178–183, 2003.

[214] F. Duong and W. Wickner, "Sec-dependent membrane protein biogenesis: SecYEG, preprotein hydrophobicity and translocation kinetics control the stop-transfer function," *EMBO J.*, vol. 17, no. 3, pp. 696–705, 1998.

[215] L. Tu, V. Santarelli, and C. Deutsch, "Truncated K^+ channel DNA sequences specifically suppress lymphocyte K^+ channel gene expression," *Biophys. J.*, vol. 68, no. 1, pp. 147–156, 1995.

[216] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, and R. MacKinnon, "X-ray structure of a voltage-dependent K^+ channel," *Nature*, vol. 423, pp. 33–41, 2003.

[217] S. B. Long, E. B. Campbell, and R. MacKinnon, "Crystal structure of a mammalian voltage-dependent *Shaker* family K^+ channel," *Science*, vol. 309, no. 5736, pp. 897–903, 2005.

[218] A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, and D. A. Doyle, "Crystal structure of the potassium channel KirBac1.1 in the closed state," *Science*, vol. 300, no. 5627, pp. 1922–1926, 2005.

[219] X. Tao, J. L. Avalos, J. Chen, and R. MacKinnon, "Crystal structure of the eukaryotic strong inward-rectifier K^+ channel Kir2.2 at 3.1 Å resolution," *Science*, vol. 326, no. 5960, pp. 1668–1674, 2009.

[220] S. Uysal, V. Vásquez, V. Tereshko, K. Esaki, F. A. Fellouse, S. S. Sidhu, S. Koide, E. Perozo, and A. Kossiakoff, "Crystal structure of full-length KcsA in its closed conformation," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 16, pp. 6644–6649, 2009.

[221] F. H. Yu and W. A. Catterall, "Overview of the voltage-gated sodium channel family," *Genome Biol.*, vol. 4, no. 3, p. 207, 2003.

[222] T. Shinoda, H. Ogawa, F. Cornelius, and C. Toyoshima, "Crystal structure of the sodium-potassium pump at 2.4 Å resolution," *Nat. Lett.*, vol. 459, pp. 446–450, 2009.

[223] R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, and R. MacKinnon, "X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity," *Nature*, vol. 415, pp. 287–294, 2002.

[224] A. Karlin, "Emerging structure of the nicotinic acetylcholine receptors," *Nat. Rev. Neurosci.*, vol. 3, pp. 102–114, 2002.

[225] P. H. Celie, I. E. Kasheverov, D. Y. Mordvintsev, R. C. Hogg, P. van Nierop, R. van Elk, S. E. van Rossum-Fikkert, M. N. Zhmak, D. Bertrand, V. Tsetlin, T. K. Sixma, and A. B. Smit, "Crystal structure of nicotinic acetylcholine receptor homolog AChBP in complex with an alpha-conotoxin PnIA variant," *Nat. Struct. Mol. Biol.*, vol. 12, no. 7, pp. 582–588, 2005.

[226] G. Chang, R. H. Spencer, A. T. Lee, M. T. Barclay, and D. C. Rees, "Structure of the MscL homolog from mycobacterium tuberculosis: a gated mechanosensitive ion channel," *Science*, vol. 282, no. 5397, pp. 2220–2226, 1998.

[227] R. B. Bass, P. Strop, M. Barclay, and D. C. Rees, "Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel," *Science*, vol. 298, no. 5598, pp. 1582–1587, 2002.

[228] S. Pegan, C. Arrabit, W. Zhou, W. Kwiatkowski, A. Collins, P. A. Slesinger, and S. Choe, "Cytoplasmic domain structures of Kir2.1 and Kir3.1 show sites for modulating gating and rectification," *Nat. Neurosci.*, vol. 8, no. 3, pp. 279–287, 2005.

[229] W. Han, S. Nattel, T. Noguchi, and A. Shrier, "C-terminal domain of Kv4.2 and associated KChIP2 interactions regulate functional expression and gating of Kv4.2," *J. Biol. Chem.*, vol. 281, no. 37, pp. 27 134–27 144, 2006.

[230] V. R. Schack, J. P. Morth, M. S. Toustrup-Jensen, A. N. Anthonisen, P. Nissen, J. P. Andersen, and B. Vilsen, "Identification and function of a cytoplasmic K^+ site of the Na^+, K^+-ATPase," *J. Biol. Chem.*, vol. 283, no. 41, pp. 27 982–27 990, 2008.

[231] F. Potet, B. Chagot, M. Anghelescu, P. C. Viswanathan, S. Z. Stepanovic, S. Kupershmidt, W. J. Chazin, and J. R. Balser, "Functional interactions between distinct sodium channel cytoplasmic domains through the action of calmodulin," *J. Biol. Chem.*, vol. 284, no. 13, pp. 8846–8854, 2009.

[232] D. Wray, "Structure and function of ion channels," *Eur. Biophys. J.*, vol. 38, no. 3, pp. 271–272, 2009.

[233] L. M. Sharkey, X. Cheng, V. Drews, D. A. Buchner, J. M. Jones, M. J. Justice, S. G. Waxman, S. D. Dib-Hajj, and M. H. Meisler, "The ataxia3 mutation in the N-terminal cytoplasmic domain of sodium channel Na(v)1.6 disrupts intracellular trafficking," *J. Neurosci.*, vol. 29, no. 9, pp. 2733–2741, 2009.

[234] T. Aoki, M. Hirano, Y. Takeuchi, T. Kobayashi, T. Yanagida, and T. Ide, "Rearrangements in the KcsA cytoplasmic domain underlie its gating," *J. Biol. Chem.*, vol. 285, no. 6, pp. 3777–3783, 2010.

[235] M. D. Edwards, I. R. Booth, and S. Miller, "Gating the bacterial mechanosensitive channels: MscS a new paradigm?" *Curr. Opin. Microbiol.*, vol. 7, no. 2, pp. 163–167, 2004.

[236] P. Koprowski and A. Kubalski, "C termini of the *Escherichia coli* mechanosensitive ion channel (MscS) move apart upon the channel opening," *J. Biol. Chem.*, vol. 278, no. 13, pp. 11 237–11 245, 2003.

[237] W. Grajkowski, A. Kubalski, and P. Koprowski, "Surface changes of the mechanosensitive channel MscS upon its activation, inactivation, and closing," *Biophys. J.*, vol. 88, no. 4, pp. 3050–3059, 2005.

[238] S. Miller, W. Bartlett, S. Chandrasekaran, S. Simpson, M. Edwards, and I. R. Booth, "Domain organization of the MscS mechanosensitive channel of *Escherichia coli*," *EMBO J.*, vol. 22, no. 1, pp. 36–46, 2003.

[239] U. Schumann, M. Edwards, C. Li, and I. R. Booth, "The conserved carboxy-terminus of the MscS mechanosensitive channel is not essential but increases stability and activity," *FEBS Lett.*, vol. 572, no. 1-3, pp. 233–237, 2004.

[240] H. R. Malcolm, Y. Y. Heo, D. B. Caldwell, J. K. McConnell, J. F. Hawkins, R. C. Guayasamin, D. E. Elmore, and J. A. Maurer, "Ss-bCNGa: a unique member of the bacterial cyclic nucleotide gated (bCNG) channel family that gates in response to mechanical tension," *Eur. Biophys. J.*, vol. 41, no. 12, pp. 1003–1013, 2012.

[241] G. Maksaev and E. S. Haswell, "MscS-Like10 is a stretch-activated ion channel from *Arabidopsis thaliana* with a preference for anions," *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 46, pp. 19 015–19 020, 2012.

[242] X. Zhang, J. Wang, Y. Feng, J. Ge, W. Li, W. Sun, I. Iscla, J. Yu, P. Blount, Y. Li, and M. Yang, "Structure and molecular mechanism of an anion-selective mechanosensitive channel of small conductance," *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 44, pp. 18 180–18 185, 2012.

[243] B. Martinac, M. Buechner, A. H. Delcour, J. Adler, and C. Kung, "Pressure-sensitive ion channel in *Escherichia coli*," *Proc. Natl. Acad. Sci. USA*, vol. 84, pp. 2297–2301, 1987.

[244] S. Sukharev, "Purification of the small mechanosensitive channel of *Escherichia coli* (MscS): the subunit structure, conduction, and gating characteristics in liposomes," *Biophys. J.*, vol. 83, no. 1, pp. 290–298, 2002.

[245] B. Akitake, A. Anishkin, and S. Sukharev, "The "dashpot" mechanism of stretch-dependent gating in MscS," *J. Gen. Physiol.*, vol. 125, no. 2, pp. 143–154, 2005.

[246] E. L. Cussler, *Diffusion: Mass Transfer in Fluid Systems.* Cambridge University Press, 1997.

[247] L. G. Longsworth, "Diffusion measurements, at 25°, of aqueous solutions of amino acids, peptides and sugars," *J. Am. Chem. Soc.*, vol. 75, no. 22, pp. 5705–5709, 1953.

[248] S. G. Schultz, N. L. Wilson, and W. Epstein, "Cation transport in *Escherichia coli*. II. Intracellular chloride concentration," *J. Gen. Physiol.*, vol. 46, no. 1, pp. 159–166, 1962.

[249] T. Ogahara, M. Ohno, M. Takayama, K. Igarashi, and H. Kobayashi, "Accumulation of glutamate by osmotically stressed *Escherichia coli* is dependent on pH," *J. Bacteriol.*, vol. 177, no. 20, pp. 5987–5990, 1995.

[250] M. Sotomayor, T. A. van der Straaten, U. Ravaioli, and K. Schulten, "Electrostatic properties of the mechanosensitive channel of small conductance MscS," *Biophys. J.*, vol. 90, no. 10, pp. 3496–3510, 2006.

[251] Y.-H. Li and S. Gregory, "Diffusion of ions in sea-water and in deep-sea sediments," *Geochim. Cosmochim. Acta*, vol. 38, no. 5, pp. 703–714, 1974.

[252] H. Tal-Ezer and R. Kosloff, "An accurate and efficient scheme for propagating the time dependent Schrödinger equation," *J. Chem. Phys.*, vol. 81, no. 9, pp. 3967–3971, 1984.